

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-348061

(43)Date of publication of application : 15.12.2000

(51)Int.Cl.

G06F 17/30  
G06F 12/00

(21)Application number : 11-162990

(71)Applicant : NIPPON TELEGR & TELEPH CORP  
<NTT>

(22)Date of filing : 09.06.1999

(72)Inventor : IIZUKA YUICHI  
TSUNAKAWA MITSUAKI  
NAGAMATSU HISAHIRO  
HOSHINO TAKASHI  
MACHIARA HIROKI

(30)Priority

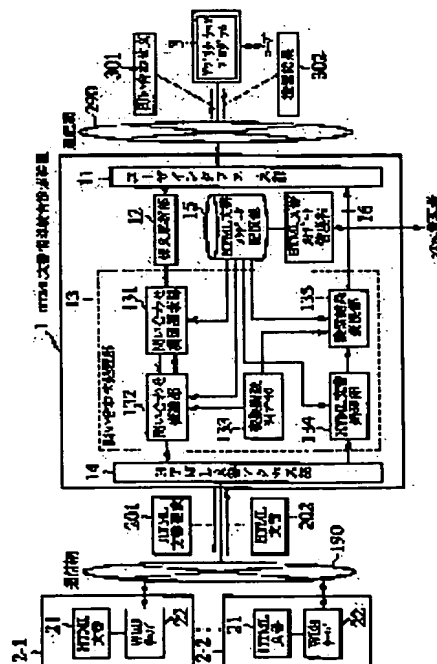
Priority number : 10162648 Priority date : 10.06.1998 Priority country : JP  
10219365 03.08.1998  
11096183 02.04.1999

JP

JP

**(54) SEMI-STRUCTURED DOCUMENT INFORMATION INTEGRATING RETRIEVAL DEVICE, SEMI-STRUCTURED DOCUMENT INFORMATION EXTRACTING DEVICE, ITS METHOD AND RECORDING MEDIUM FOR STORING ITS PROGRAM**

(57)Abstract:

**PROBLEM TO BE SOLVED:** To realize batch item unit retrieval concerning internal information over a plurality of semi-structured documents which are scattered on an open network.**SOLUTION:** An integrating retrieval device is constituted by providing a document position locating part 131 for obtaining the position of the semi-structured documents based on an input inquiry, an inquiry converting part 132 for converting an inquiry into the expression form of an item corresponding to a retrieval item in the semi-structured document, a document retrieving part 14 for obtaining the semi-structure documents by the converted inquiry, a document processing part 134 for extracting item data based on document structure information for division at every item to be extracted, selecting the extracted item data based on attribute information for retrieving the item by condition and adopting it as a retrieval result and a retrieval result converting part 135 for converting the retrieval result into the expression form of the item which is defined at every user.

## LEGAL STATUS

[Date of request for examination]	09.06.1999
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	3160265
[Date of registration]	16.02.2001
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号  
特開2000-348061  
(P2000-348061A)

(43) 公開日 平成12年12月15日 (2000. 12. 15)

(51) Int.Cl. <sup>7</sup>	識別記号	F I	テーマコード* (参考)	
G 0 6 F 17/30		C 0 6 F 15/40	3 4 0	5 B 0 7 0
12/00	5 4 6	12/00	5 4 6 R	5 B 0 8 2
		15/40	3 1 0 F	
		15/419	3 2 0	

審査請求 有 請求項の数37 O L (全 51 頁)

(21) 出願番号 特願平11-162990  
(22) 出願日 平成11年6月9日 (1999. 6. 9)  
(31) 優先権主張番号 特願平10-162648  
(32) 優先日 平成10年6月10日 (1998. 6. 10)  
(33) 優先権主張国 日本 (J P)  
(31) 優先権主張番号 特願平10-219365  
(32) 優先日 平成10年8月3日 (1998. 8. 3)  
(33) 優先権主張国 日本 (J P)  
(31) 優先権主張番号 特願平11-96183  
(32) 優先日 平成11年4月2日 (1999. 4. 2)  
(33) 優先権主張国 日本 (J P)

(71) 出願人 000004226  
日本電信電話株式会社  
東京都千代田区大手町二丁目3番1号  
(72) 発明者 飯塚 裕一  
東京都新宿区西新宿三丁目19番2号 日本  
電信電話株式会社内  
(72) 発明者 網川 光明  
東京都新宿区西新宿三丁目19番2号 日本  
電信電話株式会社内  
(74) 代理人 100083806  
弁理士 三好 秀和 (外1名)

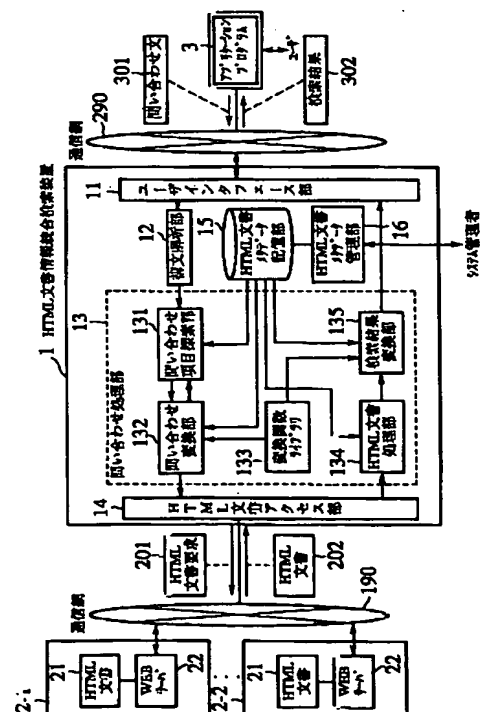
最終頁に続く

(54) 【発明の名称】 半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびにそのプログラムを格納する記録媒体

(57) 【要約】

【課題】 オープンなネットワーク上に散在する複数の半構造化文書に跨って内在する情報への一括した項目単位の検索を実現する。

【解決手段】 入力問い合わせに基づき半構造化文書の所在を得る文書所在探索部131と、問い合わせを半構造化文書中の検索項目に対応する項目の表現形式に変換する問い合わせ変換部132と、変換された問い合わせにより半構造化文書を取得する文書検索部14と、半構造化文書から、抽出すべき項目ごとに区切るための文書構造情報に基づいて、項目データを抽出し、項目を条件検索するための属性情報に基づいて前記抽出された項目データを選択して検索結果とする文書処理部134と、検索結果を、各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部135とを具備して統合検索装置を構成する。



## 【特許請求の範囲】

【請求項1】 オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索装置であって、

半構造化文書ごとに該半構造化文書中から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報を記憶する記憶部と、

入力された問い合わせから、前記メタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得る検索部と、

ユーザーごとに所定の単一フォーマットで前記検索結果を出力する出力部とを具備することを特徴とする半構造化文書情報統合検索装置。

【請求項2】 オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索装置であって、

オープンネットワーク上での半構造化文書の所在を示す所在情報と、前記半構造化文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報と、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶部と、

検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を前記所在情報から得る文書所在探索部と、

入力された前記問い合わせを、必要に応じ、前記表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換する問い合わせ変換部と、

前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索部と、

取得された各半構造化文書から、前記文書構造情報に基づいて、項目データを抽出し、必要に応じて前記検索条件を用い、前記属性情報に基づいて前記抽出された項目データを選択して検索結果とする文書処理部と、

前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置。

【請求項3】 上記半構造化文書情報統合検索装置は、さらに、半構造化文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の所定の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部を具備し、前記文書処理部は、前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する

前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とすることを特徴とする請求項2に記載の半構造化文書情報統合検索装置。

【請求項4】 前記文書処理部は、前記検索結果を、表形式に整形することを特徴とする請求項3に記載の半構造化文書情報統合検索装置。

【請求項5】 前記文書処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較することを特徴とする請求項3に記載の半構造化文書情報統合検索装置。

【請求項6】 前記テンプレートは、半構造化文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項3に記載の半構造化文書情報統合検索装置。

【請求項7】 前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、

前記文書処理部は、前記取得された半構造化文書をスキャンして、該半構造化文書の前記部分構造と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項3に記載の半構造化文書情報統合検索装置。

【請求項8】 前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項3に記載の半構造化文書情報統合検索装置。

【請求項9】 オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索装置であって、

オープンネットワーク上でのサーチエンジンの所在を示す所在情報と、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報と、HTML文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報と、ユーザーの項目の表現形式と各HTML文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶部と、

検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を前記所在情報から得る文書所在

探索部と、

前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件を満たす検索条件が指定されたサーチエンジンを、前記入力必須項目情報から得て、検索対象サーチエンジンとして選択するサーチエンジン選択部と、

前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定部と、

前記変換された問い合わせ群のそれぞれを、必要に応じ、前記表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換部と、

前記変換された問い合わせを前記得られた所在に送信して、HTML文書を取得する文書検索部と、

各サーチエンジンから取得されたHTML文書からなる第1の検索結果から、前記文書構造情報に基づいて、項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用い、前記属性情報に基づいて、前記抽出された項目データを選択して、第2の検索結果とする検索結果処理部と、

前記第2の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置。

【請求項10】 上記半構造化文書情報統合検索装置は、さらに、

HTML文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部を具備し、

前記文書処理部は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、第2の検索結果とすることを特徴とする請求項9に記載の半構造化文書情報統合検索装置。

【請求項11】 前記文書処理部は、前記検索結果を、表形式に整形することを特徴とする請求項10に記載の半構造化文書情報統合検索装置。

【請求項12】 前記文書処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較することを特徴とする請求項10に

記載の半構造化文書情報統合検索装置。

【請求項13】 前記テンプレートは、HTML文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項10に記載の半構造化文書情報統合検索装置。

【請求項14】 前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、

前記文書処理部は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項10に記載の半構造化文書情報統合検索装置。

【請求項15】 前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項10に記載の半構造化文書情報統合検索装置。

【請求項16】 オープンネットワーク上の任意のHTML文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出装置であって、

HTML文書ごとに、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部と、

取得されたHTML文書に対応するテンプレートを解析するテンプレート解析部と、

前記取得されたHTML文書をスキャンして、該HTML文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理部とを具備することを特徴とする半構造化文書情報抽出装置。

【請求項17】 前記テンプレート処理部は、前記抽出された項目データを、表形式に整形することを特徴とする請求項16に記載の半構造化文書情報抽出装置。

【請求項18】 前記テンプレート処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較することを特徴とする請求項16に記載の半構造化文書情報抽出装置。

【請求項19】 前記テンプレートは、HTML文書が複数の同一部分構造を有する場合、各部分構造ごとに抽

出テキスト形式情報が記述され、

前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項16に記載の半構造化文書情報抽出装置。

【請求項20】 前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、

前記テンプレート処理部は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項16に記載の半構造化文書情報抽出装置。

【請求項21】 前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項16に記載の半構造化文書情報抽出装置。

【請求項22】 オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索方法であって、

入力された問い合わせから、半構造化文書ごとに該半構造化文書から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得るステップと、

ユーザーごとに所定の単一フォーマットで前記検索結果を出力するステップとを含むことを特徴とする半構造化文書情報統合検索方法。

【請求項23】 オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索方法であって、

検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を、オープンネットワーク上での半構造化文書の所在を示す所在情報から得るステップと、入力された前記問い合わせを、必要に応じ、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換するステップと、前記変換された検索要求を前記得られた所在に送信して、半構造化文書を取得するステップと、

取得された各半構造化文書から、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて前記検索条件を

用い、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報に基づいて、前記抽出された項目データを選択して検索結果とするステップと、

前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換するステップとを含むことを特徴とする半構造化文書情報統合検索方法。

【請求項24】 オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索方法であって、

検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を、オープンネットワーク上でのサーチエンジンの所在を示す所在情報から得るステップと、

前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件が指定されたサーチエンジンを、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報から得て、検索対象サーチエンジンとして選択するステップと、

前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換するステップと、

前記変換された問い合わせ群のそれぞれを、必要に応じ、ユーザーの項目の表現形式と各HTML文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換するステップと、前記変換された問い合わせを前記得られた所在に送信して、HTML文書を取得するステップと、

各サーチエンジンから取得されたHTML文書からなる第1の検索結果から、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用いて項目を条件検索するための属性を規定する属性情報に基づいて前記抽出された項目データを選択して、第2の検索結果とするステップと、

前記第2の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換するステップとを含むことを特徴とする半構造化情報統合検索方法。

【請求項25】 オープンネットワーク上の任意のHTML文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出方法であって、

取得されたHTML文書に対応し、HTML文書ごとに、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを解析するステップと、  
前記取得されたHTML文書をスキャンして、該HTML文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するステップとを含むことを特徴とする半構造化文書情報抽出方法。

【請求項26】 オープンネットワーク上の複数の半構造化文書に内在する情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、  
入力された問い合わせから、半構造化文書ごとに該半構造化文書から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得る処理と、  
ユーザーごとに所定の単一フォーマットで前記検索結果を出力する処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項27】 オープンネットワーク上の複数の半構造化文書に内在する情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、  
検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を、オープンネットワーク上での半構造化文書の所在を示す所在情報から得る文書所在探索処理と、  
入力された前記問い合わせを、必要に応じ、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換する問い合わせ変換処理と、

前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索処理と、  
取得された各半構造化文書から、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて前記検索条件を用い、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報に基づいて、前記抽出された項目データを選択して検索結果とする検索結果生成処理と、  
前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換処理とを含むことを特徴とするコンピュータ読み取り

可能な記録媒体。

【請求項28】 前記検索結果生成処理は、  
前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応し、半構造化文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とすることを特徴とする請求項27に記載のコンピュータ読み取り可能な記録媒体。

【請求項29】 前記検索結果生成処理は、前記検索結果を、表形式に整形することを特徴とする請求項28に記載のコンピュータ読み取り可能な記録媒体。

【請求項30】 前記検索結果生成処理は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較することを特徴とする請求項28に記載のコンピュータ読み取り可能な記録媒体。

【請求項31】 前記テンプレートは、半構造化文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、  
前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項28に記載のコンピュータ読み取り可能な記録媒体。

【請求項32】 前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、

前記検索結果生成処理は、前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項28に記載のコンピュータ読み取り可能な記録媒体。

【請求項33】 前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項28に記載のコンピュータ読み取り可能な記録媒体。

【請求項34】 オープンネットワーク上の複数のサーチエンジンを介して情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、

検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を、オープンネットワーク上でのサーチエンジンの所在を示す所在情報から得る文書所在

探索処理と、

前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件が指定されたサーチエンジンを、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報から得て、検索対象サーチエンジンとして選択するサーチエンジン選択処理と、

前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定処理と、

前記変換された問い合わせ群のそれぞれを、必要に応じて、ユーザーの項目の表現形式と各HTML文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換処理と、

前記変換された問い合わせを前記得られた所在に送信して、HTML文書を取得する文書検索処理と、

各サーチエンジンから取得されたHTML文書からなる第1の検索結果から、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用いて項目を条件検索するための属性を規定する属性情報に基づいて前記抽出された項目データを選択して、第2の検索結果とする検索結果生成処理と、

前記第2の検索結果を、必要に応じて、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項35】 前記検索結果生成処理は、

前記取得されたHTML文書をスキャンして、該HTML文書と、該HTML文書に対応し、HTML文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とすることを特徴とする請求項34に記載のコンピュータ読み取り可能な記録媒体。

【請求項36】 前記検索結果生成処理は、前記検索結果を、表形式に整形することを特徴とする請求項35に記載のコンピュータ読み取り可能な記録媒体。

【請求項37】 前記検索結果生成処理は、前記テンプレート中の前記抽出テキスト形式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML

文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較することを特徴とする請求項35に記載のコンピュータ読み取り可能な記録媒体。

【請求項38】 前記テンプレートは、HTML文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項35に記載のコンピュータ読み取り可能な記録媒体。

【請求項39】 前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、

前記検索結果生成処理は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項35に記載のコンピュータ読み取り可能な記録媒体。

【請求項40】 前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項35に記載のコンピュータ読み取り可能な記録媒体。

【請求項41】 オープンネットワーク上の任意のHTML文書に内在する情報から項目ごとに情報を抽出する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、

取得されたHTML文書に対応し、HTML文書ごとに、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを解析するテンプレート解析処理と、

前記取得されたHTML文書をスキャンして、該HTML文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致する項目の項目データを抽出する項目データ抽出処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項42】 前記項目データ抽出処理は、前記抽出された項目データを、表形式に整形することを特徴とする請求項41に記載のコンピュータ読み取り可能な記録媒体。

【請求項43】 前記項目データ抽出処理は、前記テンプレート中の前記抽出テキスト形式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書



と前記テンプレートとを比較することを特徴とする請求項41に記載のコンピュータ読み取り可能な記録媒体。

【請求項44】 前記テンプレートは、HTML文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記項目データ抽出処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項41に記載のコンピュータ読み取り可能な記録媒体。

【請求項45】 前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、

前記項目データ抽出処理は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項41に記載のコンピュータ読み取り可能な記録媒体。

【請求項46】 前記テンプレートは、HTML文書が異なる項目を有する複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記項目データ抽出処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項41に記載のコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびにそのプログラムを格納する記録媒体に関する。特に、オープンなネットワーク上に散在する複数の半構造化文書や複数のサーチエンジンが存在する環境において、これら半構造化文書の文書構造・表現形式・構成要素や、これらサーチエンジンの入力フォーム等の相違に拘わらず、各半構造化文書の所在情報・文書構造情報・項目情報・表現形式情報等を管理することによって、各半構造化文書に跨って内在する情報への、統一インターフェースによる一括したアイテムベースの統合的検索方式を実現する半構造化文書情報統合検索技術に関する。

【0002】

【従来の技術】近年、パソコンの高機能化および低価格化、ネットワーク技術の向上、ネットワーク・プロバイダの普及と低価格化等により、インターネットに代表されるオープンネットワークが普及している。このオープンネットワークの普及に伴い、多くの情報プロバイダがコンテンツ作成を容易に実現するハイパーテキストの記述言語であるHTML (Hyper Text Markup Language) を用い、オープンネットワーク上で多種多様な情報発信を行うようになってきた。これら情報プロバイダは、情

報コンシューマが爆発的に増加したのに伴い、急激に増加してきた。情報プロバイダが増加することにより、様々な種類の情報がネットワーク中に蓄積されてきたため、これらの情報の中から、いかに情報コンシューマが要求する情報を効率よく探索するかが大きな課題となっている。

【0003】情報コンシューマは、これらの複数の情報源に対して興味対象の情報を一括して横断的に検索したいという要請を持つ。しかし、各HTML文書の構造、表現形式、各HTML文書の検索方式などは互いに異なるため、異なる情報源を横断的に検索することは困難である。

【0004】ネットワーク上に散在するHTML文書の検索に関し、一般にサーチエンジンと呼ばれる情報検索装置が普及している。ここで、サーチエンジンとは、フォーム入力によりある情報を検索するシステムの総称である。図51は、従来技術におけるURLサーチエンジンによる情報検索方式を説明する図である。ここで、URLサーチエンジンとは、キーワードや条件入力による問い合わせに対して、URLを検索結果として返す情報検索装置をいう。例えば「予算10万円以下のPCが欲しい」という要求を満たすため、ユーザーはまずURLサーチエンジンに対してキーワード検索を行う。図52は従来技術における各URLサーチエンジンの構成を示す図である。ネットワークに散在するHTML文書検索用のキーワードと文書の所在を示すURLが予めURLサーチエンジン900にキーワードインデックス910として登録されている。検索処理部930は、指定されたキーワードによりキーワードインデックス910を探索し、指定されたキーワードやその類義語を含むHTML文書の所在を示すURLのリストや概要を検索結果としてユーザーに返す。図51に戻り、ユーザーは得られたURLのHTML文書に個別にアクセスして所望する情報を人手で探索していた。即ち、HTML文書に内在する情報を検索する場合、所在が既知でないHTML文書から所望する情報を得るためには、ユーザーはまず全文検索によりHTML文書の所在を探索し、得られた所在リストの複数のHTML文書の内容閲覧を繰り返すため、所望する情報を取得するまでに多くの時間と労力を要していた。さらに、この所望する情報が複数のHTML文書に散在する場合、これらを横断的に検索することは困難であった。

【0005】この従来の方式では、指定されたキーワードやその類義語を含むHTML文書の所在情報の検索はできるが、各HTML文書に内在する情報に対するアイテムベースの一括した検索が不可能であった。また検索結果に対する条件指定（日付によるフィルタリング等）も不可能であった。さらに、各HTML文書への検索インターフェースを入力フォームとして提供しているURLサーチエンジンを利用する場合、ユーザーがURLサ

ーチエンジンごとに個別のフォーム入力インターフェースを意識する必要があり、かつURLサーチエンジンごとに個別にアクセスする必要があった。

【0006】具体的には、例えば、オープンネットワーク上で、オンライン商取引を行うオンラインショップのHTML文書は、取り扱い商品に関する情報（例えば、商品名や価格など）を1つの意味のあるデータ群として、表形式や箇条書きの形式でリスト記述する場合が非常に多い。これらオンラインショップのHTML文書に内在する情報を横断的に検索することに対する需要が高まっている。この横断的検索として、例えば“指定の商品を最安値で販売しているショップの検索”等がある。従来これらオンラインショップのHTML文書から横断的検索を行うためには、ユーザは欲しい商品の名前、メーカー名、商品種別等をキーワードとして、図51の検索方式によりHTML文書の所在情報を取得し、得られた所在のHTML文書に1つずつアクセスし、所望の商品の有無を確認する必要があった。しかし図51の検索方式は、文書の構成要素を考慮しない全文検索であるため、全く関係のないHTML文書の所在まで大量に検索してしまい、これら大量のHTML文書の中から人手で所望の商品情報を探索するのに多くの時間と労力を要していた。

【0007】このように従来の検索方式では、HTML文書中の情報を項目別に収集することができなかった。即ち“商品情報を記載した表を内包するHTML文書”に対しては“商品価格”“商品イメージ”“メーカー名”等、“店舗情報が箇条書きで記載されているHTML文書”に対しては“店舗名”“電話番号”“住所”等の項目別に情報を抽出することが困難であった。また、HTML文書からの検索結果に対して日付によるフィルタリング処理などの条件指定を行うこともできなかった。

【0008】これら項目別に管理されている情報を抽出するために、文書内部の構造や文書間の関連を独自のモデルにマッピングすることにより、仮想的なデータベースを作成する従来技術がある。この従来技術の1つの例は、N.Ashish, C.A.Knoblock, “Semi-automatic wrapper generation for internet information sources”, Proceedings of cooperative information systems, 1997. に開示されている。この技術は、HTML文書中で特定のタグ（TITLEタグ、H1タグ等）や、特定のフォントタグの内容（大きさ、色、太字・イタリック等の書体等）を持つ箇所を意味のある情報ととらえ、これらの情報を自動的に抽出するための技術である。この技術は、1つの情報の最小のまとまりが1つのHTML文書に記述され、これらHTML文書が同じ形式で記述された複数のHTML文書を対象としている。この技術は、例えば、地域ごとの気象情報が異なるHTML文書に記述されている場合に有効である。

【0009】しかし、この技術は、1つのHTML文書

に表形式や箇条書きの形式でデータ群をリスト記述することは考慮されていないため、上記のケースには適用できない。

【0010】従来技術の他の例は、J.Hammer, H.Garcia-Molina, J.Cho, R.Araha, A.Crespo, “Extracting semistructured information from the web”, Workshop on management of semistructured data, 1997. に開示されている。この技術は、OEMという独自のデータモデルを用いて下層のデータベースを構築し、このデータベースと様々な情報源の対応を管理することにより、複数の異種情報源の統合的な検索を実現する技術である。この対応管理のため、この技術はHTML文書に対してはHTMLタグ記述に依存したテンプレートファイルを用いる方式を採用している。

【0011】しかし、この技術は、HTML文書に変更が生じると仮想のデータベースに影響が及び、仮想のデータベースに変更が生じるとアプリケーションに影響が及ぶため、システムの運用、保守に多大な労力が必要であった。

【0012】さらに、オンラインショップの取扱商品情報等のためのHTML記述には、標準化された形式がないため、各HTML文書に以下の差異が生じている。

【0013】第1に、ショップにより各HTML文書の文書構造が異なる。例えば、ショップAの取扱商品はTABLEタグで記述される表形式で提示されたり、ショップBの取扱商品はULタグで記述される箇条書きで提示されたりしている。

【0014】第2に、HTML文書上の同一の取扱商品に関する情報の表現形式が異なる。例えば、価格を表す表現形式では、円、千円、万円、\$等の単位の違いや、全角、半角等の表記の違いがある。

【0015】第3に、HTML文書の同じ情報を表すデータの構成要素が異なる。例えば、商品名を示すデータの構成要素は、商品名のみの記述、商品名と型番を併せて記述、メーカー名と商品名と型番を併せて記述、等の違いがある。従来の検索方式で取得したHTML文書から所望の情報を得るため、ユーザは、これらの相違する情報を並べて比較する必要がある。これらの情報の中から所望の商品情報を探索するのに、多くの時間と労力を要していた。

【0016】さらに、複数のサーチエンジンを用いてオープンネットワーク上の情報を検索する場合、これらのサーチエンジンにはそれぞれ取り扱う情報の種類等の差異があるため、状況に応じてユーザが使い分けが必要がある。換言するとユーザは各サーチエンジンに検索要求を発行する際に、サーチエンジン固有の所在情報、検索インタフェースを意識する必要があった。

【0017】このため、第1に、ユーザによるサーチエンジンの所在情報の管理が困難であった。ユーザはサーチエンジンの所在情報を、ブックマーク等を用い個人で

管理する必要があるため、特にモバイル環境下など自端末以外の環境での検索が困難であった。

【0018】第2に、各種サーチエンジンの普及に伴う入力フォームの提供する検索インタフェースの不統一性が生じた。各サーチエンジンの普及に伴い、入力フォームが乱立している。しかし、入力フォームの構造は統一されていないため、ユーザーはサーチエンジン毎に固有の操作体系、操作手順を把握する必要がある。またユーザーは、ある検索項目の処理にどのサーチエンジンが有効であるかを把握することができない。かつ得られたHTML文書中の情報を条件処理することができない。

【0019】第3に、サーチエンジンへの検索効率の悪化が生じた。上述したようにユーザーは、所望の情報を得るまで手でサーチエンジン毎に検索を行うため、検索回数が増加し、効率が非常に悪い。

【0020】第4に、各サーチエンジンからの検索結果の項目、表現形式、文字コードなどのフォーマットが不統一であるため、ユーザーが検索結果を比較するのが困難である。

【0021】上記の各サーチエンジンの異種性を解消するため、サーチエンジンの一種であるURLサーチエンジンの共通な検索インタフェースを作成し、当該検索インタフェースと個々のURLサーチエンジンの検索インタフェースの対応を管理し、共通検索インタフェースに対する検索要求を個々の検索エンジンの検索要求に変換／実行する従来技術が、Jumon World Seek, "http://member.nifty.ne.jp/jumon" に開示されている。この技術は、共通検索インタフェースがテキストボックス1つから構成されるURLサーチエンジンを提供する。しかし、一般にURLサーチエンジンだけではなく多種多様なサーチエンジンが存在し、これらの横断的な検索を実現するためには、以下の問題点があった。

【0022】(1) 複数の入力項目の考慮が必要。

【0023】最もシンプルな入力フォーム構成では、入力項目は検索する用語を入力するテキストボックス(キーワード入力部)のみであるが、キーワードとともに他の検索条件(エリア、業種等)について同時に入力し、絞り込み検索を行うことについて配慮されている場合もある。この場合、HTMLは項目を有さない半構造化文書であるため、従来技術はシステムとして複数の入力項目をサポートできず、絞り込み検索はできなかった。

【0024】(2) 使用される入力フォームの多様性への対応が必要。

【0025】サーチエンジンで通常用いられるテキスト入力用の入力フォームのオブジェクトには、テキストボックス、複数項目中1項目を選択するラジオボタン、複数項目中で任意の複数項目を選択するセレクトボックスまたはチェックボックスなど要求条件を適切に入力するためのオブジェクトが複数存在する場合がある。この場合、従来技術ではシステムとしてテキストボックス以外

のオブジェクトをサポートしていないため、対応することができなかった。

【0026】(3) さらに、複数のサーチエンジンにわたる共通検索インタフェースを用いる場合、この共通検索インタフェースの修正時にアプリケーションを再構築する必要がある。

【0027】共通検索インタフェースに対してサーチエンジンの追加／修正／削除を行う際に、共通検索インタフェースの修正が必要になり、対応するアプリケーションを再構築しなければならない。

【0028】すなわち、従来技術においては、多種多様なサーチエンジンを取り込むことができず、システム構築／維持管理に多くの時間と労力が必要であった。

【0029】

【発明が解決しようとする課題】本発明は、上記の問題点を解決するためになされたものである。

【0030】そして、その目的とするところは、オープンなネットワークに散在する複数のHTML文書に内在する情報の文書構造、表現形式、構成要素などが互いに異なっている、これら文書を跨った情報検索を実現し、このHTML記述上の差異をユーザーごとの統一形式に変換した検索結果を返却することのできる、半構造化文書情報統合検索体系を提供することにある。

【0031】本発明の他の目的は、オープンなネットワークに複数のサーチエンジンが存在する環境において各サーチエンジン固有の入力フォームのオブジェクトを個別に管理することにより複数のサーチエンジンの異種性を解消し、ユーザーの検索要求に対して各サーチエンジン固有の検索要求を生成して検索を実行することのできる、半構造化文書情報統合検索体系を提供することにある。

【0032】本発明の他の目的は、HTML文書の所在情報、HTML文書に内在する文書の構造情報、各構成要素の属性情報をHTML文書ごとに管理することにより、所在、文書構造、属性が互いに異なる任意のHTML文書から情報を項目別に抽出することのできる半構造化文書情報統合検索体系を提供することにある。

【0033】

【課題を解決するための手段】本発明の特徴は、オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索装置であって、半構造化文書ごとに該半構造化文書中から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報を記憶する記憶部と、入力された問い合わせから、前記メタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得る検索部と、ユーザーごとに所定の単一フォーマットで前記検索結果を出力する出力部とを具備することを特徴とする半構造化文書情報統合検索装置を提供する点にある。

【0034】また、本発明の他の特徴は、オープンネッ

トワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索装置であって、オープンネットワーク上での半構造化文書の所在を示す所在情報と、前記半構造化文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報と、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶部と、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を前記所在情報から得る文書所在探索部と、入力された前記問い合わせを、必要に応じ、前記表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換する問い合わせ変換部と、前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索部と、取得された各半構造化文書から、前記文書構造情報に基づいて、項目データを抽出し、必要に応じて前記検索条件を用い、前記属性情報に基づいて前記抽出された項目データを選択して検索結果とする文書処理部と、前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置が提供される。

【0035】また、本発明の他の特徴によれば、上記半構造化文書情報統合検索装置は、さらに、半構造化文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の所定の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部を具備し、前記文書処理部は、前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とする点にある。

【0036】また、本発明の他の特徴は、前記文書処理部は、前記検索結果を、表形式に整形する点にある。

【0037】また、本発明の他の特徴は、前記文書処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較する点にある。

【0038】また、本発明の他の特徴は、前記テンプレートは、半構造化文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0039】また、本発明の他の特徴は、前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記文書処理部は、前記取得された半構造化文書をスキャンして、該半構造化文書の前記部分構造と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【0040】また、本発明の他の特徴は、前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0041】また、本発明の他の特徴は、オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索装置であって、オープンネットワーク上でのサーチエンジンの所在を示す所在情報と、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報と、HTML文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報と、ユーザーの項目の表現形式と各HTML文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶部と、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を前記所在情報から得る文書所在探索部と、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件を満たす検索条件が指定されたサーチエンジンを、前記入力必須項目情報から得て、検索対象サーチエンジンとして選択するサーチエンジン選択部と、前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定部と、前記変換された問い合わせ群のそれぞれを、必要に応じ、前記表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換部と、前記変換された問い合わせを前記得られた所在に送信して、HTML文書を取得する文書検索部と、各サーチエンジンから取得されたHTML文書からなる第1の検索結果から、前記文書構造情報に基づいて、項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用い、前記属性情報に基づいて、前記抽出された項目データを選択して、第2の検索結果

とする検索結果処理部と、前記第2の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置を提供する点にある。

【0042】また、本発明の他の特徴は、上記半構造化文書情報統合検索装置は、さらに、HTML文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部を具備し、前記文書処理部は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、第2の検索結果とする点にある。

【0043】また、本発明の他の特徴は、前記文書処理部は、前記検索結果を、表形式に整形する点にある。

【0044】また、本発明の他の特徴は、前記文書処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較する点にある。

【0045】また、本発明の他の特徴は、前記テンプレートは、HTML文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0046】また、本発明の他の特徴は、前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記文書処理部は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【0047】また、本発明の他の特徴は、前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0048】また、本発明の他の特徴は、オープンネットワーク上の任意のHTML文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出装置であって、HTML文書ごとに、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、

少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部と、取得されたHTML文書に対応するテンプレートを解析するテンプレート解析部と、前記取得されたHTML文書をスキャンして、該HTML文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理部とを具備することを特徴とする半構造化文書情報抽出装置を提供する点にある。

【0049】また、本発明の他の特徴は、前記テンプレート処理部は、前記抽出された項目データを、表形式に整形する点にある。

【0050】また、本発明の他の特徴は、前記テンプレート処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較する点にある。

【0051】また、本発明の他の特徴は、前記テンプレートは、HTML文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0052】また、本発明の他の特徴は、前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記テンプレート処理部は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【0053】また、本発明の他の特徴は、前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0054】また、本発明の他の特徴は、オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索方法であって、入力された問い合わせから、半構造化文書ごとに該半構造化文書から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得るステップと、ユーザーごとに所定の単一フォーマットで前記検索結果を出力するステップとを含むことを特徴とする半構造化文書情報統合検索方法を提供する点にある。

【0055】また、本発明の他の特徴は、オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索方法であって、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を、オープンネットワーク上での半構造化文書の所在を示す所在情報から得るステップと、入力された前記問い合わせを、必要に応じ、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換するステップと、前記変換された検索要求を前記得られた所在に送信して、半構造化文書を取得するステップと、取得された各半構造化文書から、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じ前記検索条件を用い、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報に基づいて、前記抽出された項目データを選択して検索結果とするステップと、前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換するステップとを含むことを特徴とする半構造化文書情報統合検索方法を提供する点にある。

【0056】また、本発明の他の特徴は、オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索方法であって、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を、オープンネットワーク上でのサーチエンジンの所在を示す所在情報から得るステップと、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件が指定されたサーチエンジンを、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報から得て、検索対象サーチエンジンとして選択するステップと、前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換するステップと、前記変換された問い合わせ群のそれぞれを、必要に応じ、ユーザーの項目の表現形式と各HTML文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換するステップと、前記変換された問い合わせを前記得られた所在に送信して、HTML文書を取得するステップと、各サーチエンジンから取得されたHTML文書からなる第1の検索結果から、HTML文書の構造

を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用いて項目を条件検索するための属性を規定する属性情報に基づいて前記抽出された項目データを選択して、第2の検索結果とするステップと、前記第2の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換するステップとを含むことを特徴とする半構造化文書情報統合検索方法を提供する点にある。

【0057】また、本発明の他の特徴は、オープンネットワーク上の任意のHTML文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出方法であって、取得されたHTML文書に対応し、HTML文書ごとに、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを解析するステップと、前記取得されたHTML文書をスキャンして、該HTML文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するステップとを含むことを特徴とする半構造化文書情報抽出方法を提供する点にある。

【0058】また、本発明の他の特徴は、オープンネットワーク上の複数の半構造化文書に内在する情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、入力された問い合わせから、半構造化文書ごとに該半構造化文書から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得る処理と、ユーザーごとに所定の単一フォーマットで前記検索結果を出力する処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体を提供する点にある。

【0059】また、本発明の他の特徴は、オープンネットワーク上の複数の半構造化文書に内在する情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を、オープンネットワーク上での半構造化文書の所在を示す所在情報から得る文書所在探索処理と、入力された前記問い合わせを、必要に応じ、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換する問い合わせ変換処理と、前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索処理と、

取得された各半構造化文書から、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて前記検索条件を用い、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報に基づいて、前記抽出された項目データを選択して検索結果とする検索結果生成処理と、前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体を提供する点にある。

【0060】また、本発明の他の特徴は、前記検索結果生成処理は、前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応し、半構造化文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とする点にある。

【0061】また、本発明の他の特徴は、前記検索結果生成処理は、前記検索結果を、表形式に整形する点にある。

【0062】また、本発明の他の特徴は、前記検索結果生成処理は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較する点にある。

【0063】また、本発明の他の特徴は、前記テンプレートは、半構造化文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0064】また、本発明の他の特徴は、前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記検索結果生成処理は、前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【0065】また、本発明の他の特徴は、前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0066】また、本発明の他の特徴は、オープンネットワーク上の複数のサーチエンジンを通じて情報を検索

する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を、オープンネットワーク上でのサーチエンジンの所在を示す所在情報から得る文書所在探索処理と、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件が指定されたサーチエンジンを、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報から得て、検索対象サーチエンジンとして選択するサーチエンジン選択処理と、前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定処理と、前記変換された問い合わせ群のそれぞれを、必要に応じ、ユーザーの項目の表現形式と各HTML文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換処理と、前記変換された問い合わせを前記得られた所在に送信して、HTML文書を取得する文書検索処理と、各サーチエンジンから取得されたHTML文書からなる第1の検索結果から、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用いて項目を条件検索するための属性を規定する属性情報に基づいて前記抽出された項目データを選択して、第2の検索結果とする検索結果生成処理と、前記第2の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体を提供する点にある。

【0067】また、本発明の他の特徴は、前記検索結果生成処理は、前記取得されたHTML文書をスキャンして、該HTML文書と、該HTML文書に対応し、HTML文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とする点にある。

【0068】また、本発明の他の特徴は、前記検索結果生成処理は、前記検索結果を、表形式に整形する点にある。

【0069】また、本発明の他の特徴は、前記検索結果生成処理は、前記テンプレート中の前記抽出テキスト形



式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較する点にある。

【0070】また、本発明の他の特徴は、前記テンプレートは、HTML文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0071】また、本発明の他の特徴は、前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記検索結果生成処理は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【0072】また、本発明の他の特徴は、前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0073】また、本発明の他の特徴は、オープンネットワーク上の任意のHTML文書に内在する情報から項目ごとに情報を抽出する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、取得されたHTML文書に対応し、HTML文書ごとに、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを解析するテンプレート解析処理と、前記取得されたHTML文書をスキャンして、該HTML文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致する項目の項目データを抽出する項目データ抽出処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体を提供する点にある。

【0074】また、本発明の他の特徴は、前記項目データ抽出処理は、前記抽出された項目データを、表形式に整形する点にある。

【0075】また、本発明の他の特徴は、前記項目データ抽出処理は、前記テンプレート中の前記抽出テキスト形式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較する点にある。

【0076】また、本発明の他の特徴は、前記テンプレートは、HTML文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記項目データ抽出処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0077】また、本発明の他の特徴は、前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記項目データ抽出処理は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【0078】また、本発明の他の特徴は、前記テンプレートは、HTML文書が異なる項目を有する複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記項目データ抽出処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【0079】

【発明の実施の形態】以下において、図面を用いて本発明に係る半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびにそのプログラムを格納する記録媒体の実施形態が詳細に説明される。

尚、以下の実施形態において、半構造化文書とは、HTML (Hyper Text Markup Language) 文書、SGML (Standard Generalized Markup Language) 文書、XML (eXtensive Markup Language) 文書を含む。以下、特に記載しない限り、半構造化文書をHTML文書で代表させて説明する。尚、以下の実施形態は、SGML文書およびXML文書に対しても、適宜修正して適用することができる。例えば情報検索用のサーチエンジンが具備する入力フォームなどもこのHTML文書により構成されており、以下、HTML文書にはこれら入力フォームを含むものとする。また、以下の実施形態は、例えばエレクトロニック・コマース、電子図書館や電子カタログからの情報検索など、オープンネットワーク上の複数の互いに種々の相違を有する複数のHTML文書を利用するアプリケーションに広範に適用しうる。

【0080】まず、図1および図2を参照して、本発明に係る半構造化文書情報統合検索体系の原理を説明する。

【0081】図1は、本発明に係る半構造化文書情報統合検索体系を用いる場合のユーザーの処理手順のイメージを示す。本発明に係る装置は、ユーザーから入力される検索要求(例えば、「10万円以下のPCが欲しい」)に基づいて、複数のHTML文書をユーザーに替わって柔軟に検索し、検索処理結果を一括してユーザーに送信する。この検索要求は、従来の検索用キーワードではなく、検索項目とその条件からなる簡易な構文の問い合わせ文を用いることができる。即ち、「10万円以



下」などの条件を含む検索を処理することができる。

【0082】HTML文書は、RDBのような項目単位で構造化されたデータと異なり、基本的にはプレーンテキストであるが、タグなどによりある程度データが構造化されている、いわゆる半構造化データであることを特徴とする。例えばHTML文書に内在する表、リスト、箇条書きなどの意味のある1つのデータ群が、複数のHTML文書を跨って保有されていたり、逆に複数のデータ群が1つのHTML文書に内在される場合がある。同時にこれらのデータ群のうち、ある項目に対応するデータを条件処理することができない。また、HTMLにより記述された検索用の入力フォームを有するサーチエンジンでは、検索条件として指定可能なデータ項目が固定であったり、検索条件として必須のデータ項目があったりする。こうした構造を有するHTML文書に対して、本発明に係る装置は柔軟な検索要求に対応する検索結果を一括して出力する。

【0083】図2は、本発明に係る装置の原理を示す。本発明に係る装置は、事前に登録された、各HTML文書ごとの、HTML文書の所在情報、文書構造情報、表現形式などを、HTML文書メタデータとしてHTML文書メタデータ記憶部15中に管理する。この所在情報は、例えばURLである。この文書構造情報は、HTML文書中の表、リスト、箇条書きなどの部分構造に関する情報であり、これらのデータを抽出すべき各項目にマッピングするための情報である。具体的にはこの文書構造情報は、抽出すべき項目に対応するデータがタグやスラッシュなどのデリミタで区切られているという情報であり、タグやスラッシュなどのデリミタにより識別されるHTML文書データの区切られた領域を、項目に対応付けて表現形式で管理される。この表現形式は、同じ意味を表すデータが異なる表現でHTML文書内に存在する場合の、それぞれの表現形式である。

【0084】ユーザーは、本装置の問い合わせ処理部13に、問い合わせを入力する。問い合わせ処理部13は、HTML文書メタデータ記憶部15に事前に登録されている情報を参照して、HTML文書の所在、構造、表現形式を特定する。問い合わせ処理部13は、各HTML文書を取得し、特定されたこれらの情報を用いて、各HTML文書に内在する情報を抽出し、必要に応じて条件処理を行う。このため、本発明に係る装置は、各HTML文書に内在する情報の条件検索結果を、一括してかつユーザーの表現形式に適合させてユーザーに出力することができる。従って、ユーザーは、1回の検索要求で所望する情報をネットワーク上に散在するHTML文書から一括して得ることができる。このため、検索効率が向上すると同時にネットワーク上のトラフィックが低減される。

【0085】すなわち、本発明に係る装置は、第1に、オープンネットワークに接続されているHTML等の半

構造化文書の文書構造情報を管理することにより、複数のHTML文書からの項目ベースでの検索を実現する。第2に、各サーチエンジンなどのHTML文書に散在する情報を、個々のWeb情報源に対する検索方式をユーザーに意識させずに統合的な検索を実現する。

【0086】第1の実施形態図3から図14を参照して、本発明に係る半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびに半構造化文書情報統合検索プログラムおよび半構造化文書情報抽出プログラムを格納した記録媒体の第1の実施形態であるHTML文書情報統合検索装置を詳細に説明する。

【0087】第1の実施形態は、オープンなネットワークに散在する複数HTML文書に対し、各HTML文書が内在する情報の文書構造、表現形式、表などの部分構造の構成要素が互いに異なっている、各HTML文書を跨った情報検索を実現し、またそのHTML記述上の差異をユーザー毎の統一形式に変換して検索結果を一括して返却するものである。

【0088】第1の実施形態の構成の詳細な説明を行う前に、第1の実施形態で用いられる表現形式の概念および関連する用語について説明する。HTML文書が異なると、同じ意味を表す情報でも、異なった表現形式を用いていることがある。例えば、製品価格に対して、「¥1,000」、「一千元」、「1000円」と、HTML文書の記述者によって、様々な表現形式が存在する。そこで、以下の通り、用語を定義する。

【0089】・ドメインとは、1つの表現形式である。

【0090】例) 価格で、1,000円は、「円」つき表現形式で1つのドメイン。

【0091】価格で、¥1,000 は、「¥」つき表現形式で1つのドメイン。

【0092】・ドメイングループとは、同じ意味を表すドメインの集まりである。

【0093】例) 価格、年月日、等。

【0094】・ユーザ入力ドメインとは、ユーザー毎の検索要求の入力に用いるドメインである。

【0095】例) 価格は、「円」つき表現形式、年月日は西暦かつ「/」区切り表現形式。

【0096】・ユーザ出力ドメインとは、ユーザー毎に指定する検索結果に用いるドメインである。

【0097】例) 価格は、「¥」つき表現形式、年月日は年号略号かつ「.」区切り表現形式。

【0098】・ユーザドメインとは、ユーザ入力ドメインとユーザ出力ドメインの総称である。

【0099】・ローカルドメインとは、各HTML文書上のドメインである。

【0100】例) 価格は、「円」つき表現形式。

【0101】・ドメイン変換関数とは、ユーザ入力ドメインからローカルドメイン、ローカルドメインからユーザ出力ドメインへの変換を行う関数である。

【0102】なお、ユーザ入力ドメイン、ユーザ出力ドメイン、ローカルドメインが相互に異なる場合、これらの差異はドメイン変換関数を用いて解消される。

【0103】図3は、第1の実施形態に係るHTML文書情報統合検索装置の構成を示すブロック図である。HTML文書情報統合検索装置1は、ユーザーインターフェース部11と、構文解析部12と、問い合わせ処理部13と、HTML文書アクセス部14と、HTML文書メタデータ記憶部15と、HTML文書メタデータ管理部16とにより構成される。問い合わせ処理部13は、問い合わせ項目探索部131と、問い合わせ変換部132と、変換関数ライブラリ133と、HTML文書処理部134と、検索結果変換部135とを有する。

【0104】ユーザインタフェース部11は、ユーザのアプリケーションプログラム3から入力される検索項目と検索条件からなる問い合わせ文を受け付ける。構文解析部12は、ユーザインタフェース部11で受け付けた問い合わせ文の構文を解析する。問い合わせ処理部13は、各HTML文書に内在する情報から所望する項目情報を一括して検索する。問い合わせ処理部13中で、問い合わせ項目探索部131は、問い合わせ文中で指定された項目の所在を探索する。問い合わせ変換部132は、問い合わせ文のユーザ入力ドメインをローカルドメインに変換し、HTML文書アクセス部14が送出する問い合わせ文を生成する。HTML文書処理部134は、HTML文書アクセス部14が受信したHTML文書から取得した情報に対し、問い合わせ文に応じた処理（検索項目の選択、検索条件によるフィルタリング等）を行う。検索結果変換部135は、検索されたデータの表現形式をローカルドメインからユーザ出力ドメインに変換する。HTML文書アクセス部14は、オープンなネットワークに散在するHTML文書を取得し、その内在する情報を例えば表形式などの統一フォーマットに変換する。このHTML文書アクセス部14には、HTML文書21およびこのHTML文書21を管理するWEBサーバ22からなる複数のHTML文書サーバ2-1、2-2・・・が接続されている。HTML文書メタデータ記憶部15は、各HTML文書の文書構造、HTML文書に内在する情報の表現形式や項目情報などの検索対象となるHTML文書に関する情報を記憶し管理する。この項目情報として、表などの部分構造中の構成要素と抽出すべき項目が1対1対応でない場合には、抽出すべき項目に対応させて部分構造中の構成要素は複数の構成要素として管理される。HTML文書メタデータ管理部16は、HTML文書メタデータ記憶部15に対する各種情報の入力／削除／変更を行う。システム管理者は、例えばエディタなどに実装されるHTML文書メタデータ管理部16を介して、HTML文書メタデータを登録・管理する。

【0105】図4は、HTML文書メタデータ記憶部1

5が保有するテーブルの詳細を示す。HTML文書メタデータは、各HTML文書の所在情報を管理するHTML文書テーブル151と、各HTML文書に内在する構成要素の表形式に変換するための情報を管理するHTML文書-表マッピングテーブル152と、各HTML文書の項目ごとにこの項目の属性を管理するHTML文書項目テーブル153と、各ドメインの表現形式を管理するドメインテーブル154と、ユーザごとに入力ドメインおよび出力ドメインを管理するユーザドメインテーブル155と、ドメイン変換関数を管理するドメイン変換関数テーブル156とにより構成される。

【0106】次に、第1の実施形態に係るHTML文書情報統合検索装置1の処理手順の概略を説明する。第1の実施形態の処理手順は、図5に示す検索を実行する前にHTML文書メタデータ管理部16を介して管理者がメタデータの準備を行う準備フェーズと、図6に示す検索を実行する検索フェーズの2段階のフェーズがある。

【0107】図5に示す準備フェーズでは、HTML文書の所在情報をHTML文書テーブル151に定義し（ステップS100）、HTML文書に内在する情報の表形式への対応情報をHTML文書-表マッピングテーブル152に定義し（ステップS110）、HTML文書に内在する情報の項目の属性をHTML文書項目テーブル153に定義し（ステップS120）、HTML文書に内在する情報の項目のローカルドメインをドメインテーブル154に定義し（ステップS130）、ユーザ入力ドメイン、ユーザ出力ドメインをユーザドメインテーブル155に定義し（ステップS140）、必要なドメイン間の変換関数が既存か否かについてを判定し（ステップS145）、必要なドメイン変換関数が存在しない場合、新たなドメイン変換関数を作成してドメイン変換関数テーブル156に定義する（ステップS150）。

【0108】図6に示す検索フェーズでは、まず構文解析部12はユーザからの問い合わせ文を解析し、問い合わせ項目探索部131は指定された項目の所在をHTML文書テーブル151から探索し（ステップS200）、すべての検索項目に対応する項目を保持するHTML文書をHTML文書属性テーブル153から探索し（ステップS210）、問い合わせ変換部132はステップS210で探索された項目に対応するユーザ入力ドメインとユーザ出力ドメインとローカルドメインをドメインテーブル154、ユーザドメインテーブル155から取得し（ステップS220）、全項目のユーザ入力ドメインとローカルドメインが同一か否かを判定し（ステップS225）、ユーザ入力ドメインとローカルドメインとが異なる項目に対応するドメイン変換関数を取得し、当該項目をローカルドメインの表現形式に変換する（ステップS230）。HTML文書処理部134は、HTML文書アクセス部14を介して各HTML文書を

取得して検索項目に対応する項目を抽出して検索結果を得(ステップS240)、全項目のユーザ出力ドメインとローカルドメインが同一か否かを判定し(ステップS245)、検索結果変換部135はユーザ出力ドメインとローカルドメインが異なる項目に対し、ドメイン変換関数テーブル156からドメイン変換関数を取得して当該項目をユーザ出力ドメインに変換し(ステップS250)、ユーザーインターフェース部11を介して変換後の検索結果を出力する(ステップS260)。

【0109】以下、第1の実施形態の処理手順の詳細を、図7から図14を用いて具体的に説明する。

【0110】図7(A)はショップAの商品情報を示すHTML文書のWEBブラウザでの表示例であり、図8(A)はショップBの商品情報を示すHTML文書の表示例であるとする。図7(B)は図7(A)の情報を表示するためのHTML記述例であり、図10(B)は図10(A)の情報を表示するためのHTML記述例である。

【0111】HTML文書の構造を示す文書構造としては、ショップAの商品情報は内在情報の表示に表形式を用いるため、TABLEタグを使用している。ショップBの商品情報は内在情報の表示に箇条書きを用いるため、OLタグを使用している。

【0112】表現形式としては、ショップAの商品情報は価格情報として“¥”つき表現形式を使用している。ショップBの商品情報は価格情報として“円”つき表現形式を使用している。

【0113】各HTML文書の有する構成要素としては、ショップAの商品情報は商品名を、メーカー名と商品名の構成要素から構成している。ショップBの商品情報はメーカー名と商品名が分離されて構成されている。

【0114】所在情報としては、ショップAの商品情報のURLは、“http://www.shop-a.co.jp/products.html”である。ショップBの商品情報のURLは、“http://www.shop-b.co.jp/shouhin.html”である。

【0115】上記の通り、図7(A)の商品情報と図7(B)の商品情報とは、文書構造、表現形式、構成要素のすべての面で異なっている。

【0116】(1)準備フェーズ  
まず、各HTML文書の所在情報として、図9に示すように以下のページ名とURLをHTML文書テーブル151に設定する(図5のステップS100)。

【0117】(a) ショップAの商品情報

・ページ名: Shop-A

・URL: http://www.shop-a.co.jp/products.html

(b) ショップBの商品情報

・ページ名: Shop-B

・URL: http://www.shop-b.co.jp/shouhin.html

各HTML文書に内在する構成要素の表形式への対応情報として、図10に示すように以下のページ名、レコー

ド開始点、列1～列4の取り出し方をHTML文書一表マッピングテーブル152に設定する(ステップS110)。尚、ショップBの価格情報は、数字と“,”の箇所のみが取得されるよう設定している。

【0118】(a) ショップAの商品情報

・ページ名: Shop-A

・レコード開始: “<TR><TD>”で始まる行

・列1: “ショップA”固定値

・列2: レコード開始行中の1つ目の“<TD>”と1つ目の“/”の間

・列3: レコード開始行中の1つ目の“/”と1つ目の“<TD>”の間

・列4: レコード開始行中の2つ目の“<TD>”と2つ目の“</TD>”の間

(b) ショップBの商品情報

・ページ名: Shop-B

・レコード開始: “<LI>”で始まる行

・列1: “ショップB”固定値

・列2: レコード開始行中の1つ目の“<LI>”と1つ目の“/”の間

・列3: レコード開始行中の1つ目の“/”と2つ目の“/”の間

・列4: レコード開始行中の2つ目の“/”と1つ目の“円”の間

次に、HTML文書に内在する項目の属性情報として、図11に示すように、以下のページ名、対応列、列名、データ型をHTML文書項目テーブル153に設定する(ステップ120)。ここでは、価格情報のみが数値のデータ型として定義されている。このデータ型は、検索条件の処理時に数値として比較を行うために設定されている。

【0119】(a-1) ページ名Aの列1

・ページ名: Shop-A

・対応列: 列1

・列名: ショップ名

・データ型: 文字列

(a-2) ページ名Aの列2

・ページ名: Shop-A

・対応列: 列2

・列名: メーカー名

・データ型: 文字列

(a-3) ページ名Aの列3

・ページ名: Shop-A

・対応列: 列3

・列名: 商品名

・データ型: 文字列

(a-4) ページ名Aの列4

・ページ名: Shop-A

・対応列: 列4

・列名: 価格

- ・データ型：数値
- (b-1) ページ名Bの列1
  - ・ページ名：Shop-B
  - ・対応列：列1
  - ・列名：ショップ名
  - ・データ型：文字列
- (b-2) ページ名Bの列2
  - ・ページ名：Shop-B
  - ・対応列：列2
  - ・列名：メーカー名
  - ・データ型：文字列
- (b-3) ページ名Bの列3
  - ・ページ名：Shop-B
  - ・対応列：列3
  - ・列名：商品名
  - ・データ型：文字列
- (b-4) ページ名Bの列4
  - ・ページ名：Shop-B
  - ・対応列：列4
  - ・列名：価格
  - ・データ型：数値

次に、HTML文書に内在する情報の構成要素のローカルドメインを、図12に示すようにドメインテーブル154に定義する(ステップS130)。ショップAとショップBのショップ名、メーカー名、商品名については、各々任意の文字列であるため、特にローカルドメインを設定しない。一方価格については、図11の数値のデータ型の設定値を鑑み、ショップA、ショップBのローカルドメインを以下のように定義する。同時にこのローカルドメインをHTML文書項目テーブル153に登録する。

【0120】・ドメイングループ：価格

- ・ショップAのローカルドメイン：「¥」記号つき表現形式
- ・ショップBのローカルドメイン：数値と「,」からなる表現形式

次に、ユーザ毎にユーザ入力ドメインとユーザ出力ドメインを、図13に示すようにユーザドメインテーブル155に以下のように定義する(ステップS140)。ユーザAは、ショップ名、メーカー名、商品名をHTML文書の表現形式で入力してそのままの表現形式での出力を要求するため、ユーザ入力ドメインとユーザ出力ドメインは設定しない。また、ユーザAは、価格ドメイングループについて、

- ・入力：「円」記号つき表現形式
- ・出力：「円」記号つき表現形式

を用いるものとする。このドメインの登録をドメインテーブル154にし、ユーザドメインの登録をユーザドメインテーブル155にする。尚、ユーザドメインは、ユーザ入力ドメインとユーザ出力ドメインで異なっていて

もよい。

【0121】次に、ドメイン間の変換関数を、図14に示すようにドメイン変換関数テーブル156に定義する(ステップS150)。ドメインとして、数値と「,」からなる表現形式、「円」記号つき表現形式、「¥」記号つき表現形式の3種類が存在するため、ユーザ入力ドメイン-ローカルドメイン、ユーザ出力ドメイン-ローカルドメインの相互変換用に以下の関数を以下のように作成し、ドメイン変換関数テーブル156に設定する。各変換関数は変換関数ライブラリ133に格納される。

【0122】(a) 数値と「,」からなる表現形式から「円」記号つき表現形式への変換

- ・変換関数名：Num2Yen( )
- ・変換元ドメイン：数値と「,」からなる表現形式
- ・変換先ドメイン：「円」記号つき表現形式

(b) 「円」記号つき表現形式から数値と「,」からなる表現形式への変換

- ・変換関数名：Yen2Num( )
- ・変換元ドメイン：「円」記号つき表現形式
- ・変換先ドメイン：数値と「,」からなる表現形式

(c) 数値と「,」からなる表現形式から「¥」記号つき表現形式への変換

- ・変換関数名：Num2¥( )
- ・変換元ドメイン：数値と「,」からなる表現形式
- ・変換先ドメイン：「¥」記号つき表現形式

(d) 「¥」記号つき表現形式から数値と「,」からなる表現形式への変換

- ・変換関数名：¥2Num( )
- ・変換元ドメイン：「¥」記号つき表現形式
- ・変換先ドメイン：数値と「,」からなる表現形式

(e) 「円」記号つき表現形式から「¥」記号つき表現形式への変換

- ・変換関数名：Yen2¥( )
- ・変換元ドメイン：「円」記号つき表現形式
- ・変換先ドメイン：「¥」記号つき表現形式

(f) 「¥」記号つき表現形式から「円」記号つき表現形式への変換

- ・変換関数名：¥2Yen( )
- ・変換元ドメイン：「¥」記号つき表現形式
- ・変換先ドメイン：「円」記号つき表現形式

## (2) 検索フェーズ

ユーザ「ユーザA」から以下の検索項目とその条件からなる簡易な構文の問い合わせ文が発行された場合の処理を例に説明する。

【0123】検索項目：ショップ名、メーカー名、商品名、価格

検索条件：価格 < 200,000円

まず、構文解析部12はユーザからの問い合わせを解析し、問い合わせ項目探索部131は指定された項目を検索(図6のステップS200)。指定された項目は「シ

ショップ名」、「メーカー名」、「商品名」、「価格」である。各項目と一致する列名を、HTML文書項目テーブル153から探索すると、以下のレコードが得られる。

【0124】(a) ショップ名

- ・ページ名「Shop-A」の対応列「列1」でデータ型「文字列」
- ・ページ名「Shop-B」の対応列「列1」でデータ型「文字列」

(b) メーカー名

- ・ページ名「Shop-A」の対応列「列2」でデータ型「文字列」
- ・ページ名「Shop-B」の対応列「列2」でデータ型「文字列」

(c) 商品名

- ・ページ名「Shop-A」の対応列「列3」でデータ型「文字列」
- ・ページ名「Shop-B」の対応列「列3」でデータ型「文字列」

(d) 価格

- ・ページ名「Shop-A」の対応列「列4」でデータ型「数値」
- ・ページ名「Shop-B」の対応列「列4」でデータ型「数値」

次に、問い合わせ項目探索部131はすべての検索項目に対応する項目を保持するHTML文書名を探索する（ステップS210）。上記で得られた結果に対し、すべての検索項目に対応する項目を保持するHTML文書を探索すると、以下の2組が生成される。また、各組み合わせのURLをHTML文書テーブル151から取得する。

【0125】(A) 組み合わせ1

(a) 対象ページ名：Shop-A

(b) 構成要素

- ・ショップ名：対応列「列1」でデータ型「文字列」
- ・メーカー名：対応列「列2」でデータ型「文字列」
- ・商品名：対応列「列3」でデータ型「文字列」
- ・価格：対応列「列4」でデータ型「数値」

(c) URL

<http://www.shop-a.co.jp/products.html>

(B) 組み合わせ2

(a) 対象ページ名：Shop-B

(b) 構成要素

- ・ショップ名：対応列「列1」でデータ型「文字列」
- ・メーカー名：対応列「列2」でデータ型「文字列」
- ・商品名：対応列「列3」でデータ型「文字列」
- ・価格：対応列「列4」でデータ型「数値」

(c) URL

<http://www.shop-b.co.jp/shouhin.html>

次に、問い合わせ変換部132は探索した項目に対応するユーザドメインとローカルドメインを取得する（ステ

ップS220）。この探索した項目に対応するローカルドメインはHTML文書項目テーブル153を探索して得られる。ローカルドメインがある項目については、当該ローカルドメインのドメイングループをドメインテーブル154から探索し、当該ドメイングループに対するユーザドメインをユーザドメインテーブル155から取得する。結果として、以下の組み合わせを得る。

【0126】(A) 組み合わせ1

(a) 対象ページ名：Shop-A

(b) 構成要素

- ・ショップ名：ローカルドメインなし
- ・メーカー名：ローカルドメインなし
- ・商品名：ローカルドメインなし
- ・価格：ローカルドメインは「¥」記号つき表現形式
- ユーザ入力ドメインは「円」記号つき表現形式
- ユーザ出力ドメインは「円」記号つき表現形式

(B) 組み合わせ2

(a) 対象ページ名：Shop-B

(b) 構成要素

- ・ショップ名：ローカルドメインなし
- ・メーカー名：ローカルドメインなし
- ・商品名：ローカルドメインなし
- ・価格：ローカルドメインは数値と「,」からなる表現形式

ユーザ入力ドメインは「円」記号つき表現形式

ユーザ出力ドメインは「円」記号つき表現形式

次に、問い合わせ変換部132はユーザ入力ドメインとローカルドメインが異なる項目に対し、ドメイン変換関数テーブル156から、変換元ドメインと変換先ドメインの一致する変換関数名を取得し、各HTML文書のローカルドメインに変換する（ステップ230）。双方の組み合わせにおいて、価格の表現形式が、ローカルドメインとユーザ入力ドメインとで異なるので、変換元ドメインと変換先ドメインをキーに変換関数名をドメイン変換関数テーブル156から探索する。

【0127】(A) 組み合わせ1

変換元ドメイン：「円」記号つき表現形式

変換先ドメイン：「¥」記号つき表現形式

変換関数名：Yen2¥( )

(B) 組み合わせ2

変換元ドメイン：「円」記号つき表現形式

変換先ドメイン：数値と「,」からなる表現形式

変換関数名：Yen2Num( )

各々の組み合わせに対して変換関数を実行し、以下を得る。

【0128】(A) 組み合わせ1

Yen2¥(200,000円) = ¥200,000

(B) 組み合わせ2

Yen2Num(200,000 円) = 200,000

次に、問い合わせ変換部132は各HTML文書アクセ

ス部14に対する以下の検索文を生成する。

【0129】(A) 組み合わせ1

(a) 対象ページ名: Shop-A

(b) 検索要求

検索項目: ショップ名、メーカー名、商品名、価格

検索条件: 価格 < ¥200,000

(B) 組み合わせ2

(a) 対象ページ名: Shop-B

(b) 検索要求

検索項目: ショップ名、メーカー名、商品名、価格

検索条件: 価格 < 200,000

HTML文書アクセス部14はこれらの問い合わせ文により各HTML文書毎に内在する情報の検索を実行し、HTML文書を取得して検索結果を生成する(ステップS240)。HTML文書処理部134はURLのリンク先から、各HTML文書に内在する情報を、HTML文書-表マッピングテーブル152に設定された情報に基づいて取り出し、検索条件が指定されていればフィルタリングを行い、以下の検索結果を得る。

【0130】(A) 組み合わせ1

(a) 対象ページ名: Shop-A

(b) 検索結果

・ショップ名: ショップA、メーカー名: Maker A、商品名: PC1、価格: ¥170,000

・ショップ名: ショップA、メーカー名: Maker B、商品名: PC101、価格: ¥198,000

(B) 組み合わせ2

(a) 対象ページ名: Shop-B

(b) 検索結果

・ショップ名: ショップB、メーカー名: Maker A、商品名: PC1、価格: 168,000

検索結果変換部135は、ユーザ出力ドメインとローカルドメインとが異なる項目がある場合、ドメイン変換関数を取得し、当該項目をユーザ出力ドメインに変換する(ステップS250)。上記の双方の組み合わせでは、価格が、ローカルドメインとユーザ出力ドメインとで異なるので、変換元ドメインと変換先ドメインをキーに変換関数名をドメイン変換関数テーブル156から探索する。

【0131】(A) 組み合わせ1

変換元ドメイン: 「¥」記号つき表現形式

変換先ドメイン: 「円」記号つき表現形式

変換関数名: ¥2Yen( )

(B) 組み合わせ2

変換元ドメイン: 数値と “,” からなる表現形式

変換先ドメイン: 「円」記号つき表現形式

変換関数名: Num2Yen( )

各々の組み合わせに対して変換関数を実行し、以下の結果を得る。

【0132】(A) 組み合わせ1

¥2Yen (¥170,000) = 170,000円

¥2Yen (¥198,000) = 198,000円

(B) 組み合わせ2

Num2Yen(168,000) = 168,000円

最後に、ユーザーインターフェース部11は以下の検索結果をユーザーに出力する(ステップS260)。

【0133】・ショップ名: ショップA、メーカー名: Maker A、商品名: PC1、価格: 170,000円

・ショップ名: ショップA、メーカー名: Maker B、商品名: PC101、価格: 198,000円

・ショップ名: ショップB、メーカー名: Maker A、商品名: PC1、価格: 168,000円

以上説明したように、第1の実施形態は、オープンなネットワーク上の複数HTML文書に対し、各HTML文書に内在する情報に関する各種の情報をメタデータとして管理する。このため、複数のHTML文書に内在する情報に対する一括の検索が実現でき、HTML文書間の異種性による相違を解消した検索結果を生成することができる。同時に、各HTML文書に内在する情報に関する情報をHTML文書ごと個別に管理するので、HTML文書情報統合検索装置が検索対象とするHTML文書の追加、修正、削除の作業は当該HTML文書だけに関して行えば足りる。このため、等比級数的に増加するHTML文書の本装置への検索対象としての取り込みが容易となる。

【0134】また、各HTML文書からの検索結果は、項目ごとに条件処理可能な項目データとして得られるので、HTML文書処理部134は、各HTML文書の複数の検索結果を適宜マージして1つの検索結果とし、この1つの検索結果を必要に応じて条件処理することができる。

【0135】このように、第1の実施形態によれば、オープンなネットワークに散在する複数のHTML文書に対して該複数のHTML文書に内在する情報の文書構造、構成要素、表現形式等が互いに異なってもこれら複数の文書を跨った情報検索を実現し、HTML記述上の差異をユーザ毎の統一形式に変換して一括して検索結果を返却することができる。従って従来に比較して、人手による多くの時間や労力が不要となり、検索効率が画期的に向上する。第1の実施形態は、例えば「ある製品を最安値で販売している店の名前と価格を求める」というようなエレクトロニック・コマースにおける柔軟な商品情報検索に利用可能であり、公正なエレクトロニック・コマースの市場の活性化に貢献し得る。

【0136】第2の実施形態図15から図36を参照して、本発明に係る半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびにそのプログラムを格納する記録媒体の第2の実施形態であるインターネット情報統合検索装置を詳細に説明する。

【0137】第2の実施形態は、オープンなネットワー

クに複数の情報検索装置（サーチエンジン）が散在する環境で、固有の入力フォームを持つ複数のサーチエンジンに対して各サーチエンジンの文書構造、入力フォームの必須入力項目、表現形式が互いに異なっている、サーチエンジンを跨って条件指定を含む情報検索を行い、これら入力フォームの差異を解消して全サーチエンジンから検索結果を一括して取得することを実現するものである。

【0138】尚、第2の実施形態で用いられる表現形式の概念およびこれに関連する用語は、第1の実施形態と同様である。例えば、エリア名に対しても、「神奈川県」、「神奈川」と、HTML文書の記述者や検索を実行するユーザによって、様々な表現形式が存在する。

【0139】例えば、エリアについて、神奈川県は「県」つき表現形式で1つのドメインであり、ジャンルについて、中華料理は「料理」つき表現形式で1つのドメインである。ドメイングループとしては、エリア、ジャンル、等がある。あるユーザが「神奈川県」、「中華料理」と入力する場合、ユーザ入力ドメインは「県」つき表現形式であり、ジャンルは「料理」つき表現形式である。あるユーザの出力が「神奈川県」、「中華料理」である場合、ユーザ出力ドメインは「県」つき表現形式であり、ジャンルは「料理」つき表現形式である。HTML文書から抽出した検索結果が「神奈川県」である場合、ローカルドメインは「県」つき表現形式である。

【0140】尚、同一ドメイングループ内でユーザ入力ドメイン、ユーザ出力ドメイン、ローカルドメインが相互に異なる場合、第2の実施形態でも第1の実施形態同様、ドメイン変換関数を用いて、ドメイン間の差異を解消する。

【0141】図15は、第2の実施形態に係るインターネット情報統合検索装置の構成を示すブロック図である。第2の実施形態は、図3の問い合わせ処理部13を、さらに、入力必須項目探索部136と、検索パターン判定部137と、検索結果処理部138を具備する統合検索処理部130に置き換えた点において第1の実施形態の修正である。第2の実施形態に係るインターネット情報統合検索装置10は、ユーザーインターフェース部11と、構文解析部12と、統合検索処理部130と、HTML文書メタデータ記憶部15と、HTML文書メタデータ管理部16と、HTML文書アクセス部14とから構成される。第2の実施形態に係る統合検索処理部130は、問い合わせ項目探索部131と、問い合わせ変換部132と、変換関数ライブラリ133と、入力必須項目探索部136と、検索パターン判定部137と、検索結果処理部138と、検索結果変換部135とを具備する。

【0142】尚、図3と同一の符号を付した箇所は、特に断らない限り第1の実施形態と同様であり、これらの説明は省略する。図15において、ユーザーインター

フェース部11は、ユーザのアプリケーションプログラム3から入力される検索項目と検索条件からなる問い合わせ文を受け付ける。構文解析部12は、ユーザーインターフェース部11で受け付けた問い合わせ文の構文を解析する。統合検索処理部130は、各サーチエンジンにより管理されるHTML文書に内在する項目を一括して検索する。統合検索処理部130中で、問い合わせ項目探索部131は、問い合わせ文中で指定された項目の所在を探索する。入力必須項目探索部136は、各サーチエンジンの入力フォーム上のデータ項目の不足をチェックして問い合わせ先のサーチエンジンを決定する。検索パターン判定部137は、問い合わせ文に応じた最適な検索パターンを判定して、この判定結果に従い問い合わせ文を最適化する。問い合わせ変換部132は、問い合わせ文のユーザ入力ドメインをローカルドメインに変換し、HTML文書アクセス部14が送出する問い合わせ文を生成する。検索結果処理部138は、HTML文書アクセス部14が受信したHTML文書から取得した情報に対し、問い合わせ文に応じた処理（検索項目の選択、検索条件によるフィルタリング等）を行う。検索結果処理部138はまた、抽出された情報に対して検索条件によるフィルタリング処理を行うとともに、上記で決定された検索パターンに応じてサーチエンジン側で行われた条件処理を抑止する。検索結果変換部135は、検索されたデータの表現形式をローカルドメインからユーザ出力ドメインに変換する。HTML文書アクセス部14は、生成された検索文を各サーチエンジンに送信し、オープンなネットワークに散在するHTML文書をサーチエンジンを介して取得する。このHTML文書に内在する情報が第2の実施形態により例えば表形式などの統一フォーマットに変換される。このHTML文書アクセス部14には、通信網190を介してエンジン23およびデータベース24からなる複数のサーチエンジン20-1、20-2・・・が接続されている。HTML文書メタデータ記憶部150は、各サーチエンジンの所在、各サーチエンジンの有するHTML文書の文書構造、HTML文書に内在する情報の表現形式や構成要素などの各サーチエンジンに関する情報を記憶し管理する。HTML文書メタデータ管理部16は、HTML文書メタデータ記憶部150に対する各種情報の入力／削除／変更を行う。システム管理者は、例えばエディタなどに実装されるHTML文書メタデータ管理部16を介して、HTML文書メタデータを登録・管理する。

【0143】図16は、第2の実施形態に係るHTML文書メタデータ記憶部150が保有するテーブルの詳細を示す。図4に示す第1の実施形態のHTML文書メタデータ記憶部15が有する各HTML文書の所在情報を管理するHTML文書テーブル151と、各HTML文書に内在する構成要素を表形式に変換するための情報を管理するHTML文書-表マッピングテーブル152

と、各項目ごとにこの項目の属性を管理するHTML文書項目テーブル153と、各ドメインの表現形式を管理するドメインテーブル154と、ユーザーごとに入力ドメインおよび出力ドメインを管理するユーザドメインテーブル155と、ドメイン変換関数を管理するドメイン変換関数テーブル156に加え、さらに各サーチエンジンの入力フォーム中の入力必須項目を管理する入力必須項目テーブル157とにより第2の実施形態のHTML文書メタデータ記憶部150は構成される。また検索パターン判定部137は、図28に示すような内部に各サーチエンジンへの検索パターンを決定して検索文を各サーチエンジンごとに最適な問い合わせ文に変換するための検索パターンマトリックステーブルを具備する。あるいはこの検索パターンマトリックステーブルは、HTML文書メタデータ150に含まれて構成されてもよい。

【0144】次に、第2の実施形態に係るインターネット情報統合検索装置10の処理手順の詳細および各テーブルへの設定例を説明する。第2の実施形態の処理手順は、図19に示す検索を実行する前に表現形式等の準備を行う準備フェーズと、図29に示す検索を実行する検索フェーズの2段階のフェーズがある。

【0145】図17(A)、図17(B)、図17(C)に示すサーチエンジンの入力フォームが存在する場合の例で各フェーズを説明する。図18には、図17(B)のPage-Bの入力フォームに対応するHTML記述を示す。

#### 【0146】(1) 準備フェーズ

図19に示す準備フェーズではまず、HTML文書項目テーブル153を例えば図20に示すように設定する(ステップS300)。HTML文書項目テーブル153は、各サーチエンジン入力フォームの項目について、以下の項目を管理する。図20で、ページ名は各サーチエンジンの入力フォーム名を示す。対応列は、HTML文書-表マッピングテーブル152との対応付けを行う。データ項目名は、サーチエンジン入力フォームに内在する項目を示す。「項目指定可能」とは、当該項目がこのサーチエンジンの検索結果から取得できるか否かを示す。「条件指定可能」とは、当該項目がこのサーチエンジンによる検索の際に条件指定可能か否かを示す。データ型は、数値型、文字列型等のデータの処理タイプを示す。このデータ型はフィルタリング処理時のデータの評価方法として使用する。Nameタグは、選択形式となっている項目が有するNameタグを示す。ローカルドメインは、当該列が属するドメインを示す。

【0147】次に、HTML文書テーブル151を、例えば図21に示すように設定する(ステップS310)。HTML文書テーブル151は、各サーチエンジン入力フォームの所在情報として、以下の項目を管理する。図21で、ページ名は各サーチエンジンの入力フォーム名を示す。サーチエンジンURLは、各サーチエン

ジンの所在情報となるURLを示す。

【0148】次に、HTML文書-表マッピングテーブル152を、例えば図22に示すように設定する(ステップS320)。HTML文書-表マッピングテーブル152は、各サーチエンジンから返却されるHTML文書に内在する情報の表形式への対応情報として、以下の項目を管理する。図22で、ページ名は、各サーチエンジンの入力フォーム名を示す。「レコード開始」とは、取得されたHTML文書中での結果内容が開始される行をタグ情報を用いて示す。列1から列5は、取得されたHTML文書中の、検索結果とすべきデータ項目に対応する箇所をタグ情報を用いて定義する。列1から列5のそれぞれは、図20のHTML文書項目テーブル153のページ名Page\_Aの対応列「列1」から「列5」と対応する。次に、ドメインテーブル154を、例えば図23に示すように設定する(ステップS330)。ドメインテーブル154は、HTML文書項目テーブル153でローカルドメインを設定した列について、このローカルドメイン情報として、同じ意味を表すドメインの集まりであるドメイングループと、1つの表現の集まりであるドメインを管理する。

【0149】次に、ドメイン変換関数テーブル156を、例えば図24に示すように設定する(ステップS340)。ドメイン変換関数テーブル156は、ドメイン変換関数情報として、以下の項目を管理する。図24で変換関数名は、特定のドメインから特定のドメインへ変換するための関数の名前を示す。ドメイングループは同じ意味を表すドメインの集まりを示す。変換元ドメインはドメイン関数に対し、入力するドメインを示す。変換先ドメインはドメイン関数から出力されるドメインを示す。ライブラリ名はドメイン変換を実現する変換関数ライブラリ133のファイル名を示す。

【0150】次に、ユーザドメインテーブル155を、例えば図25に示すように設定する(ステップS350)。ユーザドメインテーブル155は、ユーザがドメイングループ毎に、どのような入力ドメイン、出力ドメインを指定するかを以下の項目により管理する。図25でユーザ名は、検索要求を行うユーザの名前を示す。ユーザ入力ドメインは、ユーザがあるドメイングループに対しどのようなドメインで入力するのかを示す。ユーザ出力ドメインは、ユーザがあるドメイングループから、どのようなドメインで出力されるかを示す。

【0151】次に、入力必須項目テーブル157を、例えば図26に示すように設定する(ステップS360)。サーチエンジンによっては、入力フォーム中で入力を必須とされている項目がある。入力必須項目テーブル157は、この入力必須項目を、以下の項目により管理する。図26でページ名は、各サーチエンジンの入力フォーム名を示す。入力必須項目は、サーチエンジンに対し、必ず入力する必要のある項目名を示す。



## 【0152】(2) 検索フェーズ

図29は、第2の実施形態の検索検索実行時のフローチャートを示す。

【0153】ユーザが例えば「神奈川県にある和食料理の店」の「店名」と「電話番号」について調べたい場合の第2の実施形態の検索処理を、以下のSQLのSELECT文とWHERE文のみからなる簡易な構文の問い合わせ文が入力された場合の例で説明する。

```
SELECT 店名、電話番号 WHERE エリア=" 横浜市" and ジャンル=" 和食料理"
(1-1)
```

問い合わせ項目探索部131は、図20のHTML文書項目テーブルを参照し、検索項目および検索条件項目をデータ項目名に含むサーチエンジンを探ることにより、データ項目の所在を探索する(ステップS410)。図30にこのサーチエンジン探索結果を示す。

【0156】次に、問い合わせ項目探索部131は、ステップS410の結果からHTML文書テーブル151を参照し、「店名」、「電話番号」、「エリア」、「ジャンル」のすべての項目を満たすページを特定する(ステップS420)。この時点ではPage-A、Page-B、Page-Cが検索候補サーチエンジンとなる。

【0157】入力必須項目探索部136は、入力必須項目テーブル157を参照し、各サーチエンジンの必須項目をチェックして検索候補サーチエンジンを絞り込む

```
SELECT 店名、電話番号 WHERE エリア=" 横浜市" (1-2)
```

の問い合わせ文が入力された場合には、問い合わせ項目探索部131においては、HTML文書項目テーブル153を参照することにより、Page-A、Page-B、Page-Cはいずれも項目「店名」、「電話番号」、「エリア」を含むため、検索候補サーチエンジンとされる。

【0159】次に、入力必須項目探索部136では、以下のように検索候補サーチエンジンが絞り込まれる。Page-Aは「ジャンル」を入力必須項目とする。これは、Page-Aに対する検索では「ジャンル」という項目の指定が必須であって、指定されない場合には検索できないことを意味する。問い合わせ条件(where句)には、「ジャンル」が指定されていないため、Page-Aは入力必須項目探索部136において検索対象から除外される。

【0160】Page-Cに対する検索では、「エリア」と

Page-A :

```
SELECT 店名、電話番号 WHERE エリア=" 横浜市" and ジャンル=" 和食料理"
(2-1)
```

Page-B :

```
SELECT 店名、電話番号 WHERE エリア=" 横浜市" and ジャンル=" 和食料理"
(2-2)
```

Page-C :

```
SELECT 店名、電話番号 WHERE エリア=" 横浜市" and ジャンル=" 和食料理"
(2-3)
```

次に、検索パターン判定部137は、図28の検索パタ

【0154】まず、ユーザーインターフェース部11は問い合わせ入力を受付ける(ステップS400)。「ユーザ1」が検索項目として、「店名」と「電話番号」を指定するとし、検索条件としては「エリア=横浜市」and「ジャンル=和食料理」を指定するとすると、以下の構文の問い合わせ文が入力される。

【0155】

(ステップS430)。サーチエンジンによっては、入力が必要である条件項目が存在する。このため、ステップS420で得られた所在のサーチエンジンの中で、検索条件に指定された項目以外を入力必須項目を持つサーチエンジンを除く。問い合わせ文(1-1)の条件項目が「エリア」、「ジャンル」であるのに対し、図26に示すようにPage-Aは、条件項目「ジャンル」と一致する入力必須項目「ジャンル」を含むため検索可能なエンジンであることが分かる。同様に、Page-Bも条件項目「エリア」と一致する入力必須項目「エリア」を含むため検索可能なエンジンとなる。Page-Cも条件項目と一致する入力必須項目「エリア」、「ジャンル」を含むため検索可能なエンジンとなる。

【0158】一方、例えば、

「ジャンル」の両方の指定が必須であるため、検索対象から除外される。

【0161】一方、Page-Bの入力必須項目である「エリア」は問い合わせ条件(where句)で指定されているため、Page-Bは検索対象として選択される。

【0162】他方、入力必須項目を持たないサーチエンジンに対して上記(1-2)の問い合わせを行う場合には、このサーチエンジン(ページ)は入力必須条件がないため、問い合わせ条件(where句)が指定されていても検索できる。従って、入力必須項目探索部136で検索対象サーチエンジンとして選択される。

【0163】この時点での問い合わせ文(1-1)に基づく各サーチエンジンへのSQL文はそれぞれ以下の通りである。

【0164】

ーンマトリックスを参照して検索の処理方法を決定する

(ステップS440)。ここで、この検索パターンマトリックスを説明する。図27は第2の実施形態に係るインターネット情報統合検索装置と各サーチエンジンとの簡略化した関係を示す。ユーザーから入力される問い合わせ文の処理手順には、図27中の(a)、(b)、

(c)の3つの検索パターンがある。(a)パターンは検索要求を未処理で返却する。(b)パターンは各サーチエンジンで条件処理を行う。(c)パターンは各サーチエンジンで条件処理を行ったのち、その結果を第2の実施形態に係る装置10でフィルタリング処理する。検索パターンマトリックスは、各検索文中の検索項目がそれぞれ上記3パターンのいずれに属するかを判定するために用いられる。検索パターン判定部137は、この図28に示す検索パターンマトリックスを用いて検索を実現するための戦略を決定する。図28で、検索要求の「項目」は検索すべき項目として例えばSQLのselect句で指定された項目である。検索要求の「条件」は検索要求の検索条件として例えばSQLのwhere句で指定された項目である。エンジン(サーチエンジン)の「項目」は各サーチエンジンが検索結果として返す項目である。エンジンの「条件」は例えば各サーチエンジンの入力フォームにより規定される、各エンジンが検索要求として受け付ける条件である項目である。尚、エンジンの「項目」はHTML文書項目テーブル153の「項目指定可能」の欄の値を、エンジンの「条件」はHTML文書項目テーブル153の「条件指定可能」の欄の値を示す。処理パターン中の「検索条件値をそのまま返却」とは、指定された検索項目を処理することなく条件値を返すことを示す。「情報源から返却されたものを返却」とは、指定された検索項目に対応してサーチエンジンから戻された結果を返すことを示す。「サーチエンジンで処理」とは、指定された検索条件をサーチエンジンで処理することを示す。「装置でフィルタリング」とは、指定された検索条件に対してサーチエンジンから戻された検索結果を、検索結果処理部138で条件処理することを示す。

【0165】例えば、問い合わせ文(1-1)の場合、「店名」はselect句で指定されており、where句では指定されていない。この項目「店名」は図28の「検索要求」の「項目」欄が○で「条件」欄が×の行に相当する。一方、例えば図17(A)のサーチエンジンの入力フォームpage\_Aは、図20のHTML文書項目テーブル153を参照すると、「店名」を条件として受け取り、かつ検索結果として返すことができる。このため図28のエンジンの「項目」、「条件」欄はともに○と

フィルタリング条件:「エリア」=「横浜市」

SELECT 店名、電話番号 WHERE ジャンル=「和食料理」 (3-1)

同様の手順で、Page-B、Page-Cに対する各問い合わせ文が生成される。図32は、Page-Bについて判定された処理内容を示す。図32から、「情報検索装置で処

定まる。従って、項目「店名」は図28の上から4行目のレコードに対応する。従って「店名」のPage\_Aに対する処理パターンは、エンジンから返されたデータを項目として返し、SQLで条件を指定していないため条件は処理しないことがわかる。

【0166】一方、「エリア」はselect句で指定されておらず、where句で指定されている。この項目「エリア」は図28の「検索要求」の「項目」欄が×で「条件」欄が○の行に相当する。一方、例えば図17(A)のPage\_Aは、図20のHTML文書項目テーブル153を参照すると、「エリア」を条件として受け取ることにはできないが、「エリア」を検索結果として返すことができる。このため図28のエンジンの「項目」欄は○、「条件」欄は×と定まる。従って、項目「エリア」は図28の上から8行目の行にレコードに対応する。従って「エリア」のPage\_Aに対する処理パターンは、SQLでselect句に指定がないため項目としては返さず、エンジンでは条件として処理できないため検索結果処理部138でフィルタリング処理して返すことができる。(1-1)の問い合わせ文で指定されている他の項目「電話番号」、「ジャンル」についてもPage\_Aを対象として上記の当てはめ処理を行うことで、図28から図31のマトリックスが導出される。

【0167】図31は、検索要求とPage-Aに指定可能な項目および条件項目を各データ項目毎に判定された処理内容を示す。図31から、「サーチエンジンで処理」の欄に基づき、「ジャンル」を検索条件としてPage-Aに送信すべきことがわかる。また「装置でフィルタリング」の欄に基づき、Page-Aからの検索結果を「エリア」の条件でフィルタリング処理すべきことがわかる。また「情報源から返却されたものを返却」の欄に基づき、「店名」、「電話番号」はPage-Aからの送信結果をそのまま返却すべきことがわかる。

【0168】Page\_Aに対して、問い合わせ文(1-1)により検索する場合、Page\_Aでは「店名」と「ジャンル」が条件として指定可能だが、問い合わせ文(1-1)では「ジャンル」のみ条件指定されている。このため、「ジャンル」を「和食料理」としてPage\_Aのサーチエンジンには問い合わせ、かつ検索結果処理部138でのフィルタリング処理により、「エリア」が「横浜市」である「店名」、「電話番号」のデータを選択して検索結果とする。従って、Page-Aへの検索は上記のパターンCであり、問い合わせ文(2-1)は以下のように変換される。

【0169】

理」の欄に基づき、「エリア」を検索条件としてPage-Bに送信すべきことがわかる。「装置でフィルタリング」の欄に基づき、Page-Bからの検索結果を「ジャン

ル」の条件でフィルタリング処理すべきことがわかる。  
「情報源から返却されたものを返却」の欄に基づき、  
「店名」、「電話番号」はPage-Bからの送信結果をそのまま返却すべきことがわかる。従って、Page-Bへの

フィルタリング条件：「ジャンル」＝「和食料理」

SELECT 店名、電話番号 WHERE エリア＝「横浜市」 (3-2)

図33は、Page-Cについて判定された処理内容を示す。図33から、「情報検索装置で処理」の欄に基づき、「エリア」、「ジャンル」を検索条件としてPage-Cに送信すべきことがわかる。「装置でフィルタリング」の欄に基づき、Page-Cからの検索結果にフィルタリング処理は行わないことがわかる。「情報源から返却

フィルタリング条件：なし

SELECT 店名、電話番号 WHERE エリア＝「横浜市」 and ジャンル＝「和食料理」 (3-3)

次に、問い合わせ変換部132は、検索パターン判定部137が出力する各サーチエンジンへの問い合わせ文を、各サーチエンジンのローカルドメインに適合する検索文に変換する(ステップS450)。問い合わせ変換部132は、検索条件で指定された項目に対応するサーチエンジンの項目のうち、ローカルドメインが設定されている項目のユーザ入力ドメインとローカルドメインを、HTML文書項目テーブル153およびユーザドメインテーブル155から図34に示すように取得する。ユーザ入力ドメインとローカルドメインが異なる項目について、ドメイン変換関数テーブル156から変換関数ライブラリ133中の関数情報を取得しこれらの項目をローカルドメインの表現形式に変換する。例えば、Page-Bのデータ項目名「エリア」の場合、ローカルドメインは「Page-B-City」である。このドメイングループに対するユーザ入力ドメインは、ユーザドメインテーブル155およびドメインテーブル154より、ドメインSHITSUKIである。このため、問い合わせ変換部13

Page-A :

フィルタリング条件：「エリア」＝「横浜市」

SELECT 店名、電話番号 WHERE ジャンル＝「和食料理」 (4-1=3-1)

)

Page-B :

フィルタリング条件：「ジャンル」＝「和食料理」

SELECT 店名、電話番号 WHERE エリア＝「07」 (4-2)

検索文(4-2)では、エリア＝「横浜」はエリア＝「07」に変換されている。 【0175】

Page-C :

SELECT 店名、電話番号 FROM Page-C

WHERE エリア＝「横浜市」 and ジャンル＝「1」 (4-3)

検索文(4-3)では、ジャンル＝「和食料理」はジャンル＝「1」に変換されている。

【0176】次に、HTML文書アクセス部14は、ステップS460で得られた問い合わせ文に基づいて各サ

Page-A :

検索は上記のパターンCであり、問い合わせ文(2-2)は以下のように変換される。

【0170】

されたものを返却」の欄に基づき、「店名」、「電話番号」はPage-Bからの送信結果をそのまま返却すべきことがわかる。従って、Page-Cへの検索は上記のパターンBであり、問い合わせ文(2-3)は以下のように変換される。

【0171】

2は、ドメイン変換関数テーブル156を参照し「Shi2ValueB()」を用いて「横浜市」を入力フォーム中の選択リスト中の7番目の項目であることを示す「07」に変換する。

【0172】同様に、Page-Cのデータ項目名「ジャンル」の場合、ローカルドメインは「Page-C-Dishes」である。このドメイングループに対するユーザ入力ドメインは、ユーザドメインテーブル155およびドメインテーブル154より、ドメイン「RYOURITSUKI」である。このため、問い合わせ文変換部132は、ドメイン変換関数テーブル156を参照し「Ryouri2ValueC()」を用いて「和食料理」を選択リスト中の1番目の項目であることを示す「1」に変換する。

【0173】この時点で生成されている各サーチエンジンへの問い合わせ文および検索結果処理部138でのフィルタリング条件は、以下の通りである。

【0174】

サーチエンジン固有の以下の検索要求をそれぞれ発行する。各サーチエンジンではそれぞれ検索処理が実行される(ステップS470)。

【0177】

フィルタリング条件：「エリア」＝「横浜市」

” GET http://www. Page-a.co.jp/search-shop.cgi?category=和食料理 http  
p/1.0 ” ( 5-1 )

Page-B：フィルタリング条件：「ジャンル」＝「和食料理」

” GET http://www. Page-b.co.jp/search-shop.cgi?area=07 http/1.0 ”  
( 5-2 )

Page-C：

” GET http://www. Page-c.co.jp/search-shop.cgi?area=横浜市 & categor  
y =1 http/1.0” ( 5-3 )

次に、検索結果処理部 138 は、各サーチエンジンから返却された各 HTML 文書に内在する情報を、HTML 文書-表マッピングテーブル 152 に設定された情報に基づいて抽出する（ステップ S475）。図 35（A）は Page-B の検索結果の HTML 文書のブラウザでの表

示例を示す。図 35（B）は図 35（A）の表示に対応する HTML 記述を示す。以下に各サーチエンジンから得られた検索結果を示す。

【0178】

(a) 対象ページ名：Page-A

フィルタリング条件：「エリア」＝「横浜市」

検索結果：

・店名：A1 エリア：横浜市

電話番号：(045) \*\*\*-\*\*\*\*

・店名：A2 エリア：横須賀市

電話番号：(0468) \*\*\*-\*\*\*\*

( 6-1 )

(b) 対象ページ名：Page-B

フィルタリング条件：「ジャンル」＝「和食料理」

検索結果

・店名：B1 ジャンル：和食料理

電話番号：045-\*\*\*-\*\*\*\*

・店名：B2 ジャンル：中華料理

電話番号：045-\*\*\*-\*\*\*\*

・店名：B3 ジャンル：中華料理

電話番号：045-\*\*\*-\*\*\*\*

( 6-2 )

(c) 対象ページ名：Page-C

フィルタリング条件：なし

検索結果

・店名：C1 電話番号：045-\*\*\*-\*\*\*\*

・店名：C2 電話番号：045-\*\*\*-\*\*\*\*

( 6-3 )

次に、検索結果処理部 138 は、図 28 の検索パターンマトリックスでフィルタリング処理を行うと規定されている項目について（ステップ S480Y）、各サーチエンジンからの検索結果をフィルタリング処理する（ステップ S490）。ここで、Page-A は「エリア」を評価

せず、Page-B は「ジャンル」を評価しない。従って、これらの結果から、「エリア」＝「横浜市」、「ジャンル」＝「和食料理」の条件を満たす結果を以下のように抽出する。

【0179】

(a) 対象ページ名：Page-A

フィルタリング結果

・店名：A1 電話番号：(045) \*\*\*-\*\*\*\*

( 7-1 )

(b) 対象ページ名：Page-B

フィルタリング結果

・店名：B1 電話番号：045-\*\*\*-\*\*\*\*

( 7-2 )

(c) 対象ページ名：Page-C

フィルタリング結果

・店名：C1 電話番号：045-\*\*\*-\*\*\*\*

・店名：C2 電話番号：045-\*\*\*-\*\*\*\*

( 7-3 = 6-3 )

次に、検索結果変換部135は、検索項目で指定された項目で、ローカルドメインが指定されている項目のユーザ出力ドメインとローカルドメインをHTML文書項目テーブル153、ドメインテーブル154およびユーザドメインテーブル155から図36に示すように取得する。検索結果変換部135は、ユーザ出力ドメインとローカルドメインが異なる項目に対し、ドメイン変換関数テーブル156から関数情報を取得しユーザ出力ドメインに変換する(S500)。Page-Aのデータ項目名「電話番号」の場合、ローカルドメインとユーザ出力ドメインが一致しているため、変換は行わない。一方、Page-B、Page-Cのデータ項目名「電話番号」の場合、ローカルドメインはTel-Barであるのに対し、出力ドメインはTel-Parenである。このため、検索結果変換部135は、ドメイン変換関数テーブル156を参照してBar2Paren()を用い「045-\*\*\*\*-\*\*\*\*」を「(045) \*\*\*\*-\*\*\*\*」に変換する。Page-B、Page-Cのローカルドメインは、以下のようにユーザ出力ドメインに変換される。

【0180】入力 : 「045-\*\*\*\*-\*\*\*\*」(ドメイン: Tel-Bar)

ドメイン変換関数: Bar2Paren()

出力 : 「(045) \*\*\*\*-\*\*\*\*」(ドメイン: Tel-Paren)

以上の処理により、ユーザーインターフェース部11は、統一検索結果を以下のように返却する。ユーザー側のアプリケーションプログラムでは、これらの統一検索結果を表形式などの統一フォームで表示する(ステップS510)。

【0181】

- ・店名: A1 電話番号: (045) \*\*\*\*-\*\*\*\*
- ・店名: B1 電話番号: (045) \*\*\*\*-\*\*\*\*
- ・店名: C1 電話番号: (045) \*\*\*\*-\*\*\*\*
- ・店名: C2 電話番号: (045) \*\*\*\*-\*\*\*\*

以上説明したように、第2の実施形態によれば、オープンなネットワークに散在する複数のサーチエンジンに対して検索を要求する場合、各サーチエンジン固有の入力フォームのオブジェクトを個別に管理することにより、異種の各サーチエンジンへのインターフェースの相違を解消して、複数の任意の入力項目に対応した柔軟な検索が可能となる。このため、サーチエンジンの異種性に起因する複数のサーチエンジンから返却されたHTML文書に内在する情報の文書構造、表現形式、入力フォームの差異を解消し、統一フォーマットによる検索結果の閲覧が可能となる。従って、検索効率が向上するとともに、ネットワークへの無効なトラフィックを軽減する。また、各サーチエンジンの入力フォームを個別に登録、管理するのでHTML文書メタデータの維持管理が容易に行える。

【0182】第3の実施形態図37から図50を参照し

て、本発明に係る半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびにそのプログラムを格納する記録媒体の第3の実施形態であるHTML文書情報抽出装置を詳細に説明する。

【0183】第3の実施形態は、オープンなネットワークに散在するHTML文書に対し、各HTML文書に内在する情報を、項目別に抽出する情報検索を実現する。第3の実施形態は、図3のHTML文書処理部134を、テンプレート解析部1341と、URL-テンプレート対応表1342と、テンプレート処理部1343とにより構成した点において、第1の実施形態の修正である。尚、第3の実施形態は、図3および図15の構文解析部12、問い合わせ項目探索部131、問い合わせ変換部132、HTML文書メタデータ記憶部15、150、HTML文書メタデータ管理部16等を備えることにより上記の第1または第2の実施形態と適宜組み合わせさせて実施されてもよく、また図37に示す構成により単独で実施されてもよい。

【0184】第3の実施形態は、HTML文書から情報を項目別に抽出するために必要となるHTML文書の所在情報と、HTML文書に内在するデータの文書構造情報とを、各HTML文書ごとに設定し、これらの情報を用いてHTMLに内在する情報を項目別に抽出する。具体的には所在情報は、各HTML文書のURLとして個別に管理される。PROXYの情報は、PROXY設定ファイル中のPROXYサーバ名、PROXYポート番号として管理されてもよい。文書構造情報は、HTML文書中の表、リスト、箇条書きなどの部分構造に関する情報であり、例えば抽出すべき項目がタグやスラッシュなどのデリミタで区切られているという情報である。この文書構造情報には、各項目の列、データ型などの各項目の各種属性情報を含む。この文書構造情報は、テンプレートファイル中の項目名、抽出テキスト規定部、項目名のデータ型等として管理される。このデータ型は例えば文字型、数値型などの値を持ち、各項目を処理するための付加情報として定義される。各テンプレートファイルと検索すべきHTML文書のURLまたはファイル名は、URLまたはファイル名と、テンプレートファイル名とを有するURL-テンプレート対応表により対応付けられる。また、各HTML文書は、テンプレートファイル中の抽出テキスト規定部分が規定する表形式への対応情報を参照することにより、表形式などの統一フォーマットに変換される。尚、テンプレートファイルは、図4および図16のHTML文書-表マッピングテーブル152およびHTML文書項目テーブル153に対応する。

【0185】第3の実施形態は、これらのPROXY設定ファイル、URL-テンプレート対応表、テンプレートファイルを参照し、ユーザがURLまたはファイル名を指定すると、例えばURL指定時にはPROXY設定

ファイルを参照してHTML文書を取得した後、URL-テンプレート対応表を参照して該当するテンプレートファイル名を取得し、取得されたHTML文書を先頭から順番に1行または複数行単位でスキャンして、テンプレートファイルに記述される抽出テキスト規定部分と比較した結果に基づいて項目を抽出する。この項目抽出の際には、テンプレートファイル中で次ページへのリンクの有無を検証し、次ページへのリンクが存在する場合、このリンクがなくなるまで次ページのURLもしくはファイル名を抽出してこのページの項目を抽出する処理を繰り返し行う。テンプレートファイルを参照して項目のマッチング判定を行うことにより、HTML文書に内在する情報が表形式にマッピングされて項目単位に抽出される。第3の実施形態は、この抽出された各項目のデータをテンプレートファイルで規定されるデータ型に整形し、ユーザーに項目名と整形済み項目情報を返却する。従来の技術と比較して、HTML文書中では本来文字型である文書の構成要素のデータ型を任意に規定することにより検索条件を用いて抽出された情報を条件処理できる。さらに、第1および第2の実施形態と同様、項目データの表現形式をユーザが所望する形式に加工できる。

【0186】図37は、第3の実施形態に係るHTML文書情報抽出装置の構成を示すブロック図である。第3の実施形態に係るHTML文書情報抽出装置100は、ユーザアクセス部11と、HTML文書アクセス部14と、PROXY設定ファイル141と、HTML文書処理部134と、テンプレートファイル1345と、検索結果変換部135とで構成される。HTML文書処理部134は、テンプレート解析部1341と、URL-テンプレート対応表1342と、テンプレート処理部1343とを有する。HTML文書情報抽出装置100は、ユーザーのアプリケーションプログラム3からの問い合わせ文301に基づき、PROXYサーバー2を介してHTML文書にアクセスして、あるいは直接ローカルのHTML文書にアクセスして、これらHTML文書から得られた情報をテンプレート処理して検索結果302としてユーザーに返却する。

【0187】HTML文書情報抽出装置100は、複数のHTML文書がネットワーク上に散在する環境で、HTML文書の所在、使用されるタグの種類、内包される構成要素の種類が異なっているにもかかわらず、HTML文書から項目毎に情報を抽出するのに必要となる上記の各HTML文書の所在情報、文書構造情報を各HTML文書個別に設定することにより、HTML文書からの所望する検索結果の表形式などの統一フォーマットでの抽出を実現するものである。

【0188】HTML文書情報抽出装置100のユーザアクセス部11は、ユーザからの問い合わせ文をアプリケーションプログラム3から受信し、HTML文書アクセス部14に送信する。HTML文書アクセス部14

は、ユーザアクセス部11から受信したURLまたはファイル名に基づいて必要に応じてPROXY設定ファイル141を参照して、HTML文書4-1、4-2を取得する。この取得されたHTML文書4-1、4-2をテンプレート解析部1341に送信する。HTML文書アクセス部14はまた、取得されたHTML文書がリンク情報を含む場合には、テンプレート解析部1341が抽出したリンク先URLに基づいて、必要に応じてPROXY設定ファイル141を参照して、HTML文書4-1、4-2を取得する。PROXY設定ファイル141は、図39に示すように、HTML文書4-1、4-2を取得するために必要なPROXYサーバの所在情報であるPROXYサーバ名、PROXYポート番号を規定したファイルであり、HTML文書アクセス部14により参照される。テンプレートファイル1345は、図40に示すように、HTML文書4-1、4-2から項目として抽出可能な部位および抽出項目を抽出テキスト規定部分に規定し、各抽出項目のデータ型を規定するファイルであり、テンプレート解析部1341により参照される。URL-テンプレート対応表1342は、受信したURL情報を元に、当該URLまたはファイル名がどのテンプレートと対応しているかを管理するファイルであって、テンプレート解析部1341によって参照される。テンプレート解析部1341は、URL-テンプレート対応表1342を参照して、問い合わせ文に対応するテンプレートファイル1345の名称を取得する。同時に、このテンプレートファイル名を有するテンプレートファイル1345を参照し、取得されたHTML文書の抽出可能な部位、抽出すべき項目、抽出すべき項目のデータ型等を解析、取得し、テンプレート処理部1343へ送信する。この際テンプレートファイル1345上でリンク先URLの有無も判断され、テンプレート解析部1341はリンク先が存在する場合にはHTML文書アクセス部14にリンク先URLを送信してリンク先HTML文書を取得する。テンプレート処理部1343は、テンプレート解析部1341から受信した抽出可能な部位、抽出すべき項目、抽出すべき項目のデータ型に基づいてHTML文書4-1、4-2から各項目を抽出する。検索結果変換部135は、テンプレート処理部1343から抽出されたデータおよびそのデータ型を受信し、データ型に沿った抽出データの変換処理を行う。この変換後の抽出データを検索結果302としてユーザインターフェース部11に送出する。

【0189】なお、このHTML文書情報抽出装置100は、第1および第2の実施形態と同様、CPU、メモリ、入出力装置、外部記憶装置等からなるコンピュータと、該コンピュータに読み取られた際、このコンピュータを前記各手段として機能させるためのプログラムを記憶した媒体とによって実現することもできる。

【0190】PROXYサーバ2は、HTML文書情報

抽出装置100から指定されることが可能なHTML文書取得の仲介を行うサーバであり、URLによって指定されたHTML文書4-1をHTML文書情報抽出装置100に返却する。HTML文書4-1、4-2は、オープンなネットワーク上に散在するホームページを記述したタグ付きテキストファイルである。アプリケーションプログラム3は、ユーザからのURLまたはファイル名と、少なくとも検索項目を含む問い合わせ文を受け付け、HTML文書情報抽出装置100からの受信結果をユーザに出力する。

【0191】次に、第3の実施形態に係るHTML文書情報抽出装置100の処理手順を説明する。第3の実施形態の処理手順は、図38に示す検索を実行する前に表現形式等の準備を行う準備フェーズと、図41に示す検索を実行する検索フェーズの2段階のフェーズがある。尚、図38の準備フェーズの手順は管理者が適当なエディタ等を用いて作成・設定するものであり、HTML文書情報抽出装置100自体を動作させて行うものではない。

#### 【0192】(1) 準備フェーズ

図38に示す準備フェーズでは、まず図39に示すようにPROXYサーバが必要な場合(ステップS600Y)、PROXYサーバ名、PROXYポート番号を定義してPROXY設定ファイル171が作成される(ステップS605)。次に、テンプレートファイルが作成される(ステップS610)。このテンプレートファイルには、複数のテンプレートファイル間で一意となるファイル名が与えられ、図40に例として示すように以下の情報が記述される。

#### 【0193】(a) 抽出項目

この抽出項目は、図40の「Word」キーワードに対応する。

【0194】HTML文書から抽出したい情報として、抽出すべき項目名、抽出すべき項目のデータ型、抽出すべき項目に付け加える固定値を記述する。図40でこのデータ型は、「1」が文字型を示す。尚、このデータ型には、「3」が数値型、「4」が文字列を追加する型等と所望する条件処理に応じて設定することができる。図40のテンプレートファイルには、リンク先アドレス(URLの相対パス)等が「NextURL」で始まる部分に記述されている。これらの抽出項目のデータ型及び抽出項目に付け加える固定値は、ユーザーに検索結果を返却する際に必要な情報を追加もしくは削除するために必要となる。

#### 【0195】(b) 抽出テキスト規定部分

この抽出テキスト規定部分は、図40の「HtmlTemplate」部分に対応する。

【0196】抽出対象となるWebページより、抽出したい情報を含むHTML文書のレコード分をコピーする。そのうち、取得したい情報部分を「\$抽出項目名

\$」に置き換え、残った記述のうちレコードに依存している省略可能な記述を、省略記号「…」に置き換える。

【0197】また、同一HTML文書内に異なるテーブルとして取り扱うべき情報が混在する場合、同一テーブルの最後を特定する文字列を記入する。図40では、第1、第2および第3の表の項目がそれぞれ定義されている。

【0198】さらにリンク先のURLが存在する場合、リンク先URLを特定する文字列を記入する。

【0199】次に、URL-テンプレート対応表を作成する(ステップS620)。各URLまたはファイルに対し、図41に示すように対応するテンプレートファイル名を記述したファイルを作成する。

#### 【0200】(2) 実行フェーズ

図42は、第3の実施形態が取得したHTML文書から所望する項目を抽出する実行フェーズの処理手順を示すフローチャートである。

【0201】まず、ユーザーインターフェース部11は、ユーザーがアプリケーションプログラム3に入力したURLまたはファイル名と、検索項目を含む検索文を受け付ける(ステップS700)。HTML文書アクセス部14は入力URLの場合、PROXY設定ファイル141があればそれを参照してHTML文書4-1を取得する。入力がファイル名の場合、ローカルのHTML文書が指定される。ユーザアクセス部110により受信されたURLまたはファイル名とPROXY設定ファイル141の記述内容に従って、HTML文書アクセス部14はPROXYサーバ2を介するか、直接HTML文書を取得する。また、HTML文書アクセス部14は返却結果であるHTML文書4-1を受信する(ステップS710)。

【0202】テンプレート解析部1341は、URLと対応するテンプレートファイルの有無を判定する。ユーザーインターフェース部11を介し受信したURLまたはファイル名を参照し、このURLまたはファイル名に対応するテンプレートファイル名を図41のURL-テンプレート対応表1342から探索する(ステップS720)。対応するテンプレートファイルが存在しない場合(S720N)、ユーザーインターフェース部11に対しエラーメッセージを送信する。一方存在すれば(S720Y)、テンプレート解析部1341は、取得されたHTML文書に対応するテンプレート名のテンプレートファイル1345に記述されている抽出ルールを解析し(ステップS730)、抽出に必要な情報をテンプレート処理部1343に送信する。

【0203】テンプレート処理部1343は、テンプレートファイル1345の抽出ルールを用いて、HTML文書4-1から実際に項目を抽出して表形式のデータを得る(ステップS740)。テンプレート処理部1343は、ステップS730の抽出ルール解析によりリンク

先URLの有無を判定する(ステップS750)。リンク先のURLが取得された場合(ステップS750Y)、リンク先URLをHTML文書アクセス部14に送信する。HTML文書アクセス部14により取得されたリンク先HTML文書に対してステップS730～S750の処理を行う。

【0204】検索結果変換部135は、抽出された項目の項目データを、図40のテンプレートファイル1345を参照することで、以下のデータ変換処理を行って加工する。

【0205】a)．抽出した情報をそのまま表示すべきデータ型の項目データに、変換は実施しない。

【0206】b)．固定値を代入すべきデータ型の項目データには、HTML文書中に存在しないが、項目として返却したい項目について検索結果変換処理部135が有する固定値を返却する。

【0207】c)．取得情報からカンマを削除すべきデータ型の項目データには、数値情報中からカンマを削除する。

【0208】d)．取得情報に追加すべきデータ型の項目データには、URLの相対パス等、抽出項目に対し付加すべき固定値が存在する場合、当該固定値を付加する。

【0209】上記の処理で得られるすべての検索結果は、ユーザインターフェース部11を介してアプリケーションプログラム3に送信され、表示される。

【0210】図43～図46は第3の実施形態による項目情報抽出の具体例を示すもので、図43はHTML文書のWebブラウザでの表示例、図44は図43の表示に対応するHTML記述例(但し、その一部分)である。図45は、図43、図44のHTML文書からの項目抽出を行うためのテンプレートファイル171を示すもので、各抽出項目、ここではレース名(racename)、格(grade)、競馬場(cercle)、月日(mmdd)、距離(distance)、天候・馬場(condition)、タイム(time)、勝ち馬(winhorse)、性別(sex\_age)、騎手(jockey)、調教師(teki)、リンク先(url)の各項目と、これら各項目を抽出するための抽出テキスト規定部分とが記述されている。図46は、図43、図44のHTML文書から図45のテンプレートファイル171を用いて項目抽出(検索)を行った結果の一表示例を示す。この表示例はアプリケーションプログラム3側で3つの項目(「騎手」「勝ち馬」「レース名」)を検索項目として指定または選択した場合を示す。

【0211】次に、図40、図47～図50を参照して、第3の実施形態の変形例を説明する。第3の実施形態では、図40に示すように同一HTML文書内の第1および第2の表は、同一の構成要素からなる2つの部分

構造に対応してテンプレートが定義されている。尚、ここで部分構造とは、例えば表、リスト、箇条書きなどで表現される意味のある1つのデータ群をいう。一方この変形例は、第1に同一HTML文書内の任意の項目が他の項目と異なる属性情報を含む場合にも対応できるテンプレートを用いて上記の項目抽出を行う点、第2に同一HTML文書内の異なる項目からなる複数の部分構造に対応できるテンプレートを用いて上記の項目抽出を行う点、第3にHTML文書がリンクを含む場合にも対応できるテンプレートを用いて上記の項目抽出を行う点において、第3の実施形態の変形である。

【0212】図47、図48は、店名情報を示すHTML文書のWebブラウザによる表示例を示す。図47と図48とは、それぞれ3つの表からなり、同様の文書構造を有するHTML文書である。図49は、図47の表示に対応するHTML記述を、図50は、図48の表示に対応するHTML記述を示す。図40は、図47および図48(図49および図50)から項目を抽出するためのテンプレートを示す。図40のテンプレートは、表または箇条書きなどのHTML文書中の部分構造の終端(TableEndDelimiter)、抽出項目名(Word)、抽出項目のデータ型(Word)、抽出テキスト規定部(HtmlTemplate)に関する記述からなる。例えば、HTML文書中の</TABLE>の出現を部分構造の終端とすることを、TableEndDelimiter=</TABLE>と記述する。

【0213】図49が示す<A HREF="/html\_2.html">は、図50のHTML文書へのリンクを示す。テンプレート解析部1341は、このリンク情報を解析する。テンプレート処理部1343は、このリンク情報に従い図40のテンプレートの記述(NextURL)に基づいて、図47のHTML文書のみでなく図48のHTML文書からテンプレートを参照して項目抽出を行う。

【0214】図47の表示に対応する図49のHTML記述中第1の表と第2の表とは、同一構成要素の文書構造、同一表示形式の情報を備えた2つの部分構造である。テンプレート処理部1343は、図40のテンプレートの第1および第2の部分構造(図53では表)に関する記述に基づき、同一HTML文書内の同一文書構造の複数の部分構造の項目情報を抽出する。図48の表示に対応する図50のHTML記述は図49のHTML記述と同様の文書構造を有し、図40のテンプレートにより図49のHTMLソース記述と同様に項目情報が抽出される。

【0215】尚、図49のHTMLソース記述中第1の表と第2の表とは、異なる属性(図49では表示属性)を含む2つの部分構造である。図49のHTML記述中構成要素「ジャンル」に対応するデータは、<I>と</I>で囲まれた構造のものと、そうでない構造のものがある。この「I」タグは、対応するデータをイタリック書体で表示することを示す。同様に「B」タグは、対



応するデータを太字で表示することを示す。これらの異なる属性に関する情報は、図40のテンプレート上では、同一行について2つの記述として定義されている。取得されたHTML文書がいずれかの行の記述に合致すれば、対応する項目が抽出される。図40では、上記属性に対応する記述として、省略を示すタグ「. . .」が用いられているので、任意の属性を有するデータを抽出することができる。

【0216】一方、図47の表示に対応する図49のHTMLソース記述中第1および第2の表に対し第3の表は、異なる抽出項目に対応する構成要素「評価」に対応するデータを含む部分構造である。テンプレート処理部1343は、図40の第3の表に対応する記述に基づいて、同一HTML文書内の異なる構成要素の文書構造の複数部分構造を抽出する。

【0217】以上説明したように、第3の実施形態によれば、複数の任意のHTML文書に対し、当該HTML文書が内包する情報に関する各種の情報を管理し、当該情報を用いてユーザに対し適切な情報を項目別に抽出し、表形式などの統一フォーマットにて提供することが可能となる。また、ユーザが要求する抽出対象のみを抽出テキスト規定部分に規定することにより、システム構築/維持管理が容易となる。即ち、各HTML文書が有する多種多様なインタフェースの相違に拘わらず、オープンなネットワーク上に散在する複数のHTML文書から、情報を項目別に抽出することが可能となり、抽出した情報をユーザが所望する形式により提供することが可能となる。

【0218】このように、第3の実施形態は、HTMLの構文規則に依存しないテンプレートを用いて、HTML文書から所望する項目を抽出する。即ち、タグまたはこれに準ずるデリミタ付きテキストであれば項目の抽出が可能である。また、抽出のための情報を定義するテンプレートファイルを作成するだけで、この項目の抽出を行う。テンプレートファイルは、対象となるHTML文書に基づき容易に作成可能であり、かつ視覚的にわかりやすいため、容易かつ柔軟にHTML文書に内在する情報の項目別の抽出を実現することができる。

【0219】尚、本発明は、上述した実施の形態に限定されるものではなく、その要旨を逸脱しない範囲において、種々変更することが可能である。

【0220】

【発明の効果】以上説明したように、本発明によれば、オープンなネットワークに散在する複数のHTML文書に対して該複数のHTML文書に内在する情報の文書構造、構成要素、表現形式等が互いに異なってもこれら複数の文書を跨った情報検索を実現し、HTML記述上の差異をユーザ毎の統一形式に変換して一括して検索結果を返却することができる。

【0221】さらに、各HTML文書が有する多種多様

なインタフェースの相違に拘わらず、オープンなネットワーク上に散在する複数のHTML文書から、情報を項目別に抽出することが可能となり、抽出した情報をユーザが所望する形式により提供することが可能となる。

【0222】また、オープンなネットワークに散在する複数のサーチエンジンに対して検索を要求する場合、各サーチエンジン固有の入力フォームのオブジェクトを個別に管理することにより、異種の各サーチエンジンへのインターフェースの相違を解消して、複数の任意の入力項目に対応した柔軟な検索が可能となる。

【0223】従って従来に比較して、人手による多くの時間や労力が不要となり、検索効率が画期的に向上する。

【図面の簡単な説明】

【図1】本発明に係るHTML文書情報統合検索のユーザの処理手順を説明する図である。

【図2】本発明に係るHTML文書情報統合検索装置の原理を説明する図である。

【図3】本発明の第1の実施形態に係るHTML文書情報統合検索装置の構成を示すブロック図である。

【図4】第1の実施形態に係るHTML文書メタデータ記憶部が有するテーブルの構成を説明する図である。

【図5】第1の実施形態における準備フェーズの処理手順を示すフローチャートである。

【図6】第1の実施形態における検索フェーズの処理手順を示すフローチャートである。

【図7】あるHTML文書における表示およびHTML記述の一例を示す図である。

【図8】他のHTML文書における表示およびHTML記述の一例を示す図である。

【図9】HTML文書テーブルの内容を示す図である。

【図10】図7(B)および図8(B)に対応するHTML文書-表マッピングテーブルの内容を示す図である。

【図11】図7および図8に対応するHTML文書項目テーブルの内容を示す図である。

【図12】ドメインテーブルの内容を示す図である。

【図13】ユーザードメインテーブルの内容を示す図である。

【図14】ドメイン変換関数テーブルの内容を示す図である。

【図15】本発明の第2の実施形態に係るインターネット情報統合検索装置の構成を示すブロック図である。

【図16】第2の実施形態に係るHTML文書メタデータ記憶部が有するテーブルの構成を説明する図である。

【図17】第2の実施形態で使用される各サーチエンジンの入力フォームの例を説明する図である。

【図18】図17(B)の入力フォームのHTML記述を示す図である。

【図19】第2の実施形態における準備フェーズの処理

手順を示すフローチャートである。

【図20】第2の実施形態におけるHTML文書項目テーブルの内容の一例を説明する図である。

【図21】第2の実施形態におけるHTML文書テーブルの内容の一例を説明する図である。

【図22】第2の実施形態におけるHTML文書-表マッピングテーブルの内容の一例を説明する図である。

【図23】第2の実施形態におけるドメインテーブルの内容の一例を示す図である。

【図24】第2の実施形態におけるドメイン変換関数テーブルの内容の一例を示す図である。

【図25】第2の実施形態におけるユーザードメインテーブルの内容の一例を示す図である。

【図26】第2の実施形態の入力必須項目テーブルの内容の一例を示す図である。

【図27】検索要求処理における図15の第2の実施形態に係るインターネット情報統合検索装置と各サーチエンジンとの関係を説明する図である。

【図28】第2の実施形態の検索パターンマトリックステーブルの内容を示す図である。

【図29】第2の実施形態における検索フェーズの処理手順を示すフローチャートである。

【図30】図29のステップS410で探索されたデータ項目の所在を示す図である。

【図31】図29のステップS440で得られたページAに対する検索要求の処理パターンを示す図である。

【図32】図29のステップS440で得られたページBに対する検索要求の処理パターンを示す図である。

【図33】図29のステップS440で得られたページCに対する検索要求の処理パターンを示す図である。

【図34】図29のステップS450で得られたユーザー入力ドメインとローカルドメインとの対応情報示す図である。

【図35】ページBに対する検索要求の処理結果の表示例およびHTML記述を示す図である。

【図36】図29のステップS500で得られたユーザー出力ドメインとローカルドメインとの対応情報示す図である。

【図37】本発明の第3の実施形態に係るHTML文書情報抽出装置の構成を示すブロック図である。

【図38】第3の実施形態における準備フェーズの処理手順を示すフローチャートである。

【図39】PROXY設定ファイルの内容の一例を示す図である。

【図40】第3の実施形態におけるテンプレートファイルの内容の一例を示す図である。

【図41】URL-テンプレート対応表の内容の一例を示す図である。

【図42】第3の実施形態における実行フェーズの処理手順を示すフローチャートである。

【図43】HTML文書のWebブラウザによる表示の一例を示す図である。

【図44】図43の表示に対応するHTML記述の一部を示す図である。

【図45】図43、図44に対応するテンプレートファイルの内容を示す図である。

【図46】第3の実施形態が図43のHTML文書から抽出した検索結果の表示の一例を示す図である。

【図47】第3の実施形態の変形例におけるHTML文書のWebブラウザによる表示の一例を示す図である。

【図48】図47のHTML文書からリンクされる図47の文書と同一の文書構造を有するHTML文書のWebブラウザによる表示の一例を示す図である。

【図49】図47の表示に対応するHTML記述を示す図である。

【図50】図48の表示に対応するHTML記述を示す図である。

【図51】従来のHTML文書情報検索のユーザーの処理手順を説明する図である。

【図52】従来のHTML文書情報検索の原理を説明する図である。

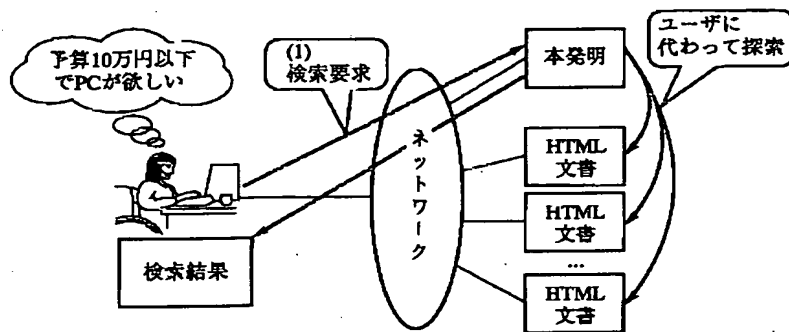
【符号の説明】

- 1 HTML文書情報統合検索装置
- 2 PROXYサーバ
- 3 アプリケーションプログラム
- 10 インターネット情報統合検索装置
- 11 ユーザーインターフェース部
- 12 構文解析部
- 13 問い合わせ処理部
- 14 HTML文書アクセス部
- 15、150 HTML文書メタデータ記憶部
- 16 HTML文書メタデータ管理部
- 4、21、202 HTML文書
- 22 Webサーバ
- 23 サーチエンジン
- 24 データベース
- 100 HTML文書情報抽出装置
- 131 問い合わせ項目探索部
- 132 問い合わせ項目変換部
- 133 変換関数ライブラリ
- 134 HTML文書処理部
- 135 検索結果変換部
- 136 入力必須項目探索部
- 137 検索パターン判定部
- 138 検索結果処理部
- 139 マトリックステ이블
- 151 HTML文書テーブル
- 152 HTML文書-表マッピングテーブル
- 153 HTML文書項目テーブル
- 154 ドメインテーブル

155 ユーザドメインテーブル  
 156 ドメイン変換関数テーブル  
 157 入力必須項目テーブル  
 190、290 通信網  
 201 HTML文書要求  
 203 検索要求  
 204 検索結果

301 問い合わせ文  
 302 検索結果  
 1341 テンプレート解析部  
 1342 URL/テンプレート対応表  
 1343 テンプレート処理部  
 1345 テンプレートファイル

【図1】



【図9】

HTML文書テーブル

ページ名	URL
Shop_A	http://www.shop_a.co.jp/products.html
Shop_B	http://www.shop_b.co.jp/shoubin.html

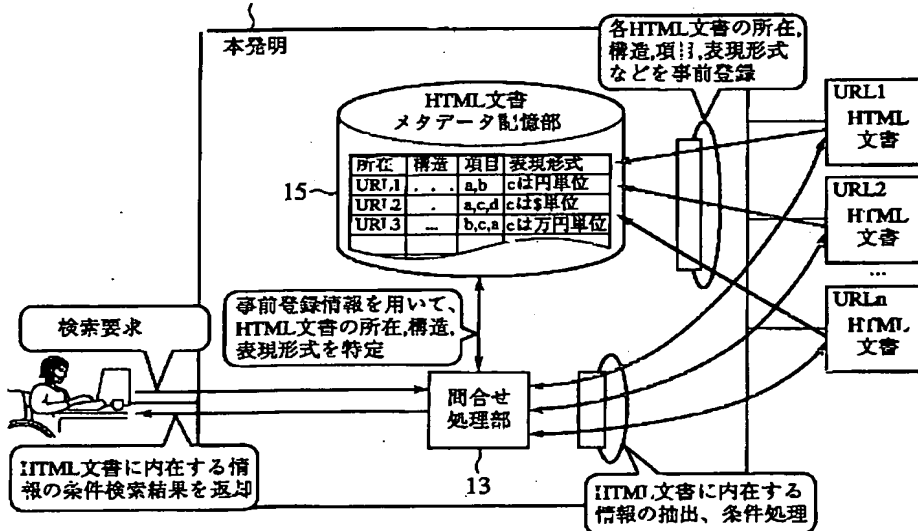
【図12】

ドメインテーブル

ドメイングループ	ユーザドメイン
価格	「¥」記号つき表現形式
価格	数値と「,」からなる表現形式
価格	「円」記号つき表現形式

【図2】

1 HTML文書情報統合検索装置



【図10】

【図13】

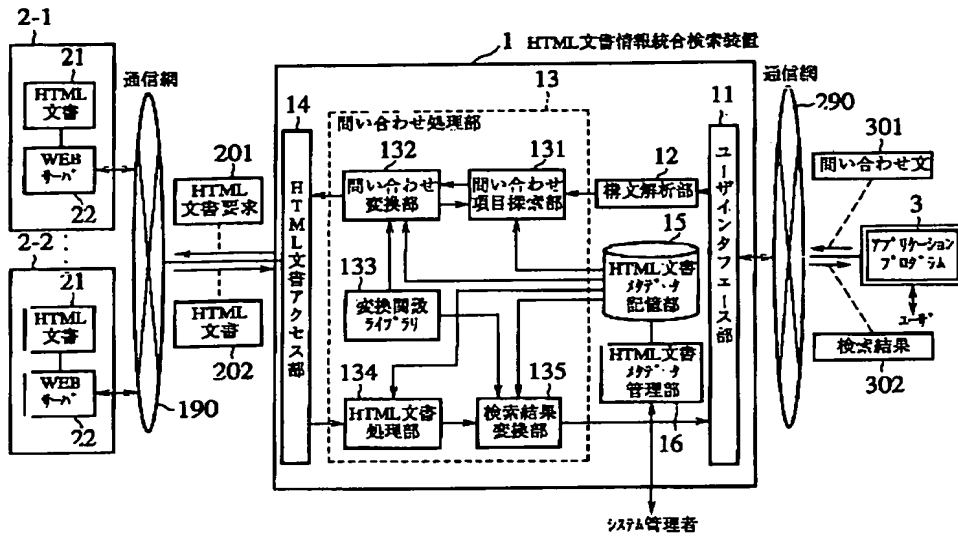
HTML-表マッピングテーブル

ページ名	ワード開始	列1	列2	列3	列4
Shop_A	<TR><TD>で始まる行	「Shop_A」固定値	ワード開始行中の1つ目の「<TD>」と1つ目の「<TD>」の間	ワード開始行中の1つ目の「<TD>」と1つ目の「<TD>」の間	ワード開始行中の2つ目の「<TD>」と2つ目の「<TD>」の間
Shop_B	<LI>で始まる行	「Shop_B」固定値	ワード開始行中の1つ目の「<LI>」と1つ目の「<LI>」の間	ワード開始行中の1つ目の「<LI>」と2つ目の「<LI>」の間	ワード開始行中の2つ目の「<LI>」と2つ目の「<LI>」の間

ユーザドメインテーブル

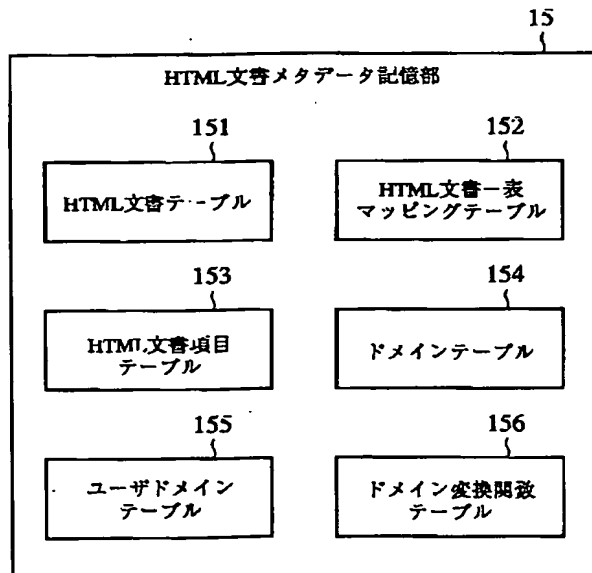
ユーザ名	ドメイングループ	ユーザ入力ドメイン	ユーザ出力ドメイン
ユーザA	価格	「円」記号つき表現形式	「円」記号つき表現形式

【図3】



【図4】

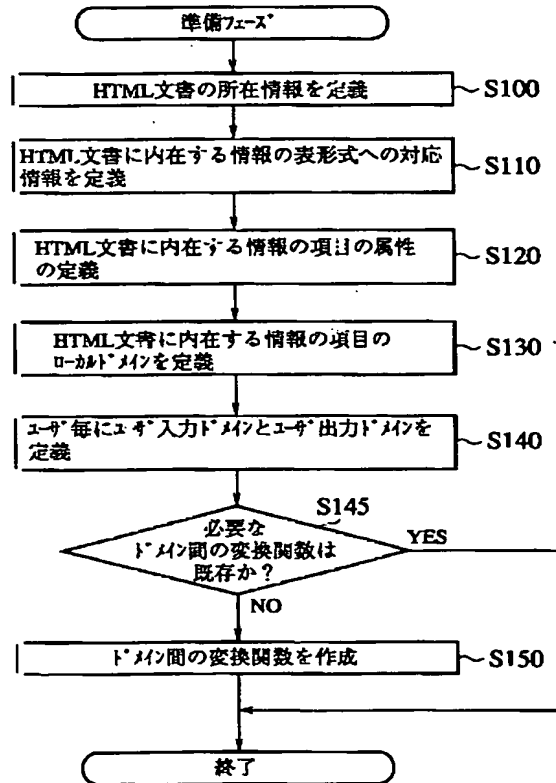
【図5】



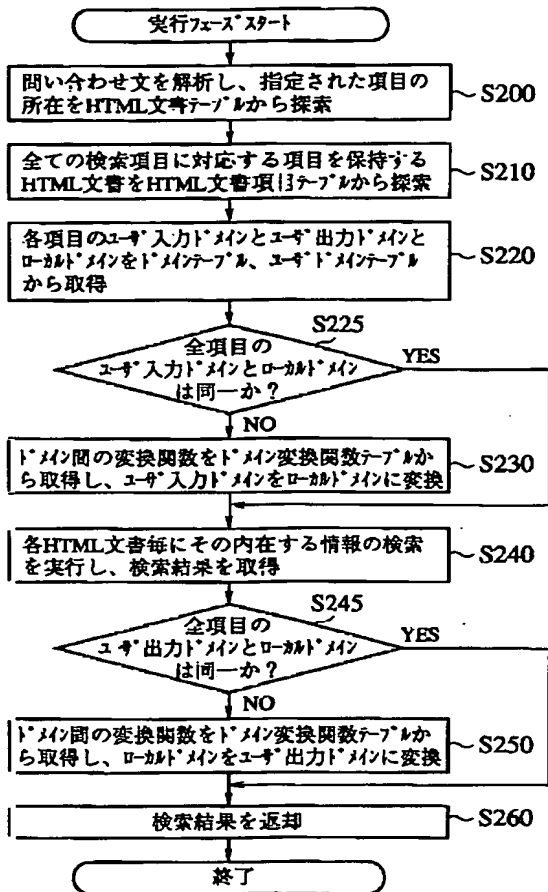
【図11】

HTML文書項目テーブル

ページ名	対応列	列名	データ型	ドメイン
Shop_A	列1	ショップ名	文字列	なし
Shop_A	列2	メーカー名	文字列	なし
Shop_A	列3	商品名	文字列	なし
Shop_A	列4	価格	数値	「¥」記号つき表現形式
Shop_B	列1	ショップ名	文字列	なし
Shop_B	列2	メーカー名	文字列	なし
Shop_B	列3	商品名	文字列	なし
Shop_B	列4	価格	数値	数値と「,」からなる表現形式



【図6】



【図8】

HTMLの内在情報の記述にOLタグを用いた例

WEBブラウザでの表示例  
 タイトル:ショップBの取扱商品情報  
 URL:http://www.shop\_b.co.jp/shouhin.html

(A)

商品情報		
メーカー名/商品名/価格		
1. Maker A/PC1/168,000円		
2. Maker B/PC101/208,000円		
3. Maker B/PC102/248,000円		

HTML文書

(B)

```

<BODY>
<H1>商品情報 </H1>
<H3>メーカー名/商品名/価格 </H3>
<OL>
<LI>Maker A/PC1/168,000円
<LI>Maker B/PC101/208,000円
<LI>Maker B/PC102/248,000円
</OL>
</BODY>
  
```

【図7】

HTMLの内在情報の記述にTABLEタグを用いた例

WEBブラウザでの表示例  
 タイトル:ショップAの取扱商品情報  
 URL:http://www.shop\_a.co.jp/products.html

(A)

商品情報		
商品名		価格
Maker A/PC1		¥170,000
Maker A/PC2		¥238,000
Maker B/PC101		¥198,000

HTML文書

(B)

```

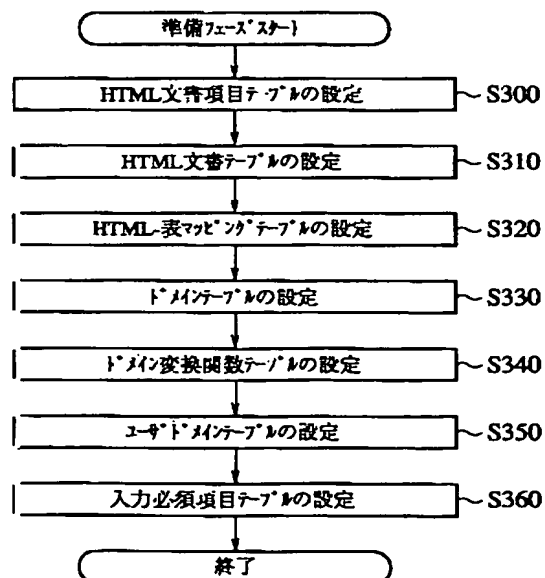
<BODY>
<H1>商品情報 </H1>
<TABLE BORDER=1>
<TR><TH>商品名</TH><TH>価格</TH></TR>
<TR><TD>Maker A/PC1</TD><TD>¥170,000</TD></TR>
<TR><TD>Maker A/PC2</TD><TD>¥238,000</TD></TR>
<TR><TD>Maker B/PC101</TD><TD>¥198,000</TD></TR>
</TABLE>
</BODY>
  
```

【図14】

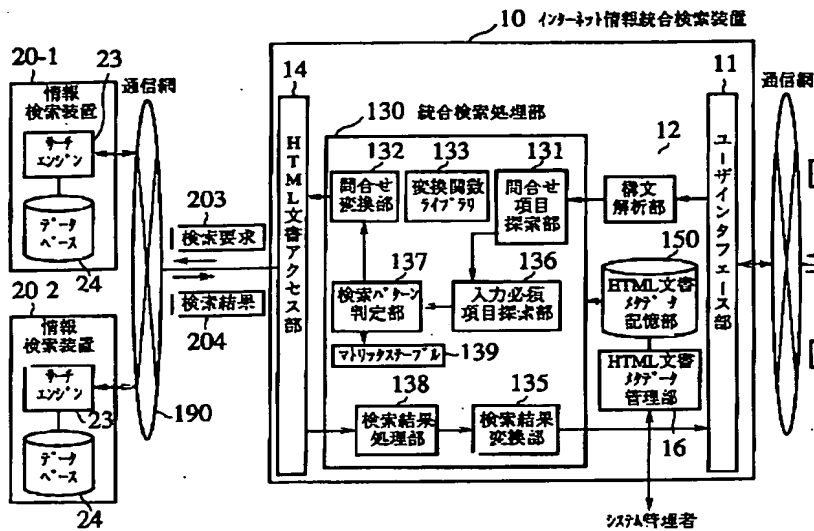
ドメイン変換関数テーブル

変換関数名	変換元ドメイン	変換先ドメイン	ファイル名
Num2Yen 0	数値と","からなる表現形式	「円」記号つき表現形式	price.dll
Yen2Num 0	「円」記号つき表現形式	数値と","からなる表現形式	price.dll
Num2Y 0	数値と","からなる表現形式	「¥」記号つき表現形式	price.dll
Y2Num 0	「¥」記号つき表現形式	数値と","からなる表現形式	price.dll
Yen2Y 0	「円」記号つき表現形式	「¥」記号つき表現形式	price.dll
Y2Yen 0	「¥」記号つき表現形式	「円」記号つき表現形式	price.dll

【図19】



【図15】

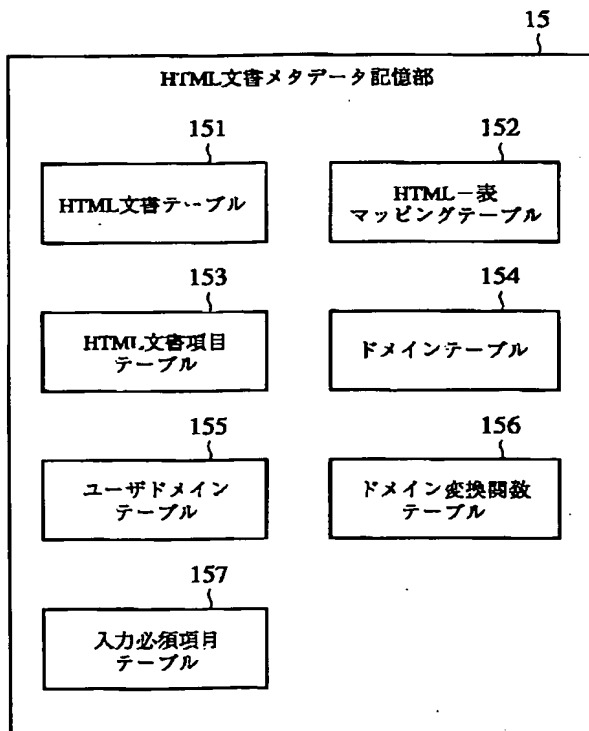


【図30】

データ項目名	ページ名テーブル
データ項目名	ページ名
店名	Page_A, Page_B, Page_C
電話番号	Page_A, Page_B, Page_C
エリア	Page_A, Page_B, Page_C
ジャンル	Page_A, Page_B, Page_C

【図16】

【図18】



Page\_BのHTML.

```

<!DOCTYPE HTML PUBLIC "-//W3C//DTD W3 HTML//EN">
<HTML>
<HEAD>
<TITLE>page_B</TITLE>
</HEAD>
<BODY>
<FORM action=http://www.page_b.co.jp/cgi-bin/search.cgi method=GET>
検索したい店名<BR><BR>
<TABLE>
<TR>
<TH>
店名
</TH>
<TH>
エリア
</TH>
<TR>
<TD>
<INPUT name=key size=30>
</TD>
<TD>
<SELECT name=area>
<OPTION value=00 SELECTED>
<OPTION value=01>横浜賀市
<OPTION value=02>藤沢市
<OPTION value=03>平塚市
<OPTION value=04>厚木市
<OPTION value=05>相模原市
<OPTION value=06>横浜原市
<OPTION value=07>横浜原市
<OPTION value=08>茅ヶ崎市
</SELECT>
</TD>
</TR>
</TABLE>
<BR>
<INPUT type=submit value=検索>
<INPUT type=reset value=クリア>
</FORM>
</BODY>
</HTML>

```

【図21】

【図34】

ページ名	情報検索装置URL
Page_A	http://www.Page_a.co.jp/search-shop.cgi
Page_B	http://www.Page_b.co.jp/search-shop.cgi
Page_C	http://www.Page_c.co.jp/search-shop.cgi

ページ名	項目	ユーザ入力ドメイン	ドキュメントドメイン
Page_B	エリア	SHITSUKI	Page_B_City
Page_C	ジャンル	RYOURITSUKI	Page_C_Dishes

【図17】

(A)

Page\_A

検索したい店

店名

ジャンル

検索 クリア

(B)

Page\_B

検索したい店

店名 エリア

検索 クリア

(C)

Page\_C

検索したい店

エリア ジャンル  
 ☐ 和食  
☐ 洋食  
☐ 中華  
☐ その他

検索 クリア

【図26】

入力必須項目テーブル	
ページ名	入力必須項目
Page_A	ジャンル
Page_B	エリア
Page_C	エリア、ジャンル

【図20】

HTML文書項目テーブル							
ページ名	対応列	項目名	項目指定可能	条件指定可能	データ型	Nameタグ	コメント
Page_A	列1	店名	○	○	文字列	shop	なし
Page_A	列2	エリア	○	-	文字列	-	なし
Page_A	列3	ジャンル	○	○	文字列	category	なし
Page_A	列4	電話番号	○	-	数値	-	Tel_Paren
Page_A	列5	URL	○	-	文字列	-	なし
Page_B	列1	店名	○	○	文字列	key	なし
Page_B	列2	エリア	○	○	文字列	area	Page_B_City
Page_B	列3	ジャンル	○	-	文字列	-	なし
Page_B	列4	電話番号	○	-	数値	-	Tel_Bar
Page_C	列1	店名	○	-	文字列	-	なし
Page_C	列2	エリア	○	○	文字列	area	なし
Page_C	列3	ジャンル	○	○	文字列	category	Page_C_Dishes
Page_C	列4	URL	○	-	文字列	-	なし
Page_C	列5	電話番号	○	-	数値	-	Tel_Bar

【図23】

ドメインテーブル		
ドメイングループ	ドメイン	説明
City	SHITSUKI	"市"つき表現形式
City	Page_B_City	Valueの値による表現形式
Dishes	RYOURITSUKI	"料理"つき表現形式
Dishes	Page_C_Dishes	Valueの値による表現形式
Tel	Tel_Bar	***-**-****からなる形式
Tel	Tel_Paren	(***)**-****からなる形式

【図39】

PROXYサーバ名	PROXYポート番号
Abc.ntt.co.jp	80

【図41】

【図22】

HTML文書一表マッピングテーブル					
ページ名	シート開始	列1	列2	列3	列4
Page_A	"<TR><TD>" で始まる行	シート開始行中の 1つ目の"<TD>" と1つ目の "</TD>"の間	シート開始行中の 2つ目の"<TD>" と2つ目の "</TD>"の間	シート開始行中の 3つ目の"<TD>" と3つ目の "</TD>"の間	シート開始行中の 4つ目の"<TD>" と4つ目の "</TD>"の間
Page_B	"<TR><TD>" で始まる行	シート開始行中の 1つ目の"<TD>" と1つ目の "</TD>"の間	シート開始行中の 2つ目の"<TD>" と2つ目の "</TD>"の間	シート開始行中の 3つ目の"<TD>" と3つ目の "</TD>"の間	シート開始行中の 4つ目の"<TD>" と4つ目の "</TD>"の間
Page_C	"<TR><TD>" で始まる行	シート開始行中の 1つ目の"<TD>" と1つ目の "</TD>"の間	シート開始行中の 2つ目の"<TD>" と2つ目の "</TD>"の間	シート開始行中の 3つ目の"<TD>" と3つ目の "</TD>"の間	シート開始行中の 4つ目の"<TD>" と4つ目の "</TD>"の間

URL/ファイル名	テンプレート名
http://www.aaa.co.jp/	aaa
http://www.bbb.co.jp/	bbb
http://www.ccc.co.jp/	ccc
http://www.ddd.co.jp/	ddd
C:\HTML\eee.htm	eee

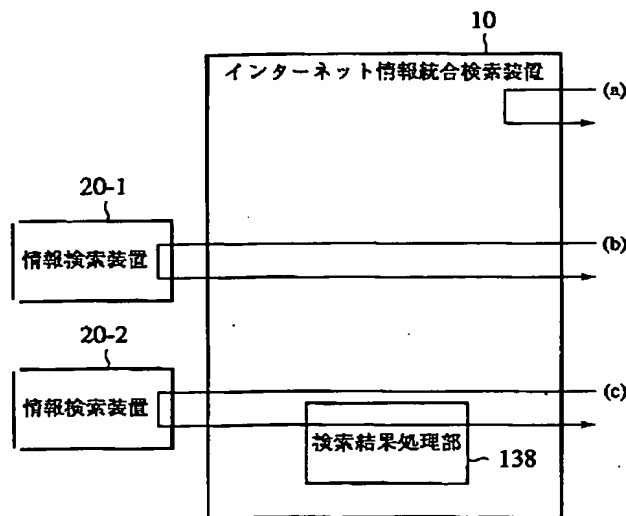
【図24】

ドメイン変換関数テーブル				
変換関数名	ドメイングループ	変換元ドメイン	変換先ドメイン	出力形式
Shi2ValueB ()	Area	SHITSUKI	Page_B_City	select.dll
Rymuri2ValueC ()	Category	RYOURITSUKI	Page_C_Dishes	select.dll
Bar2Paren ()	Tel	Tel_Bar	Tel_Paren	select.dll

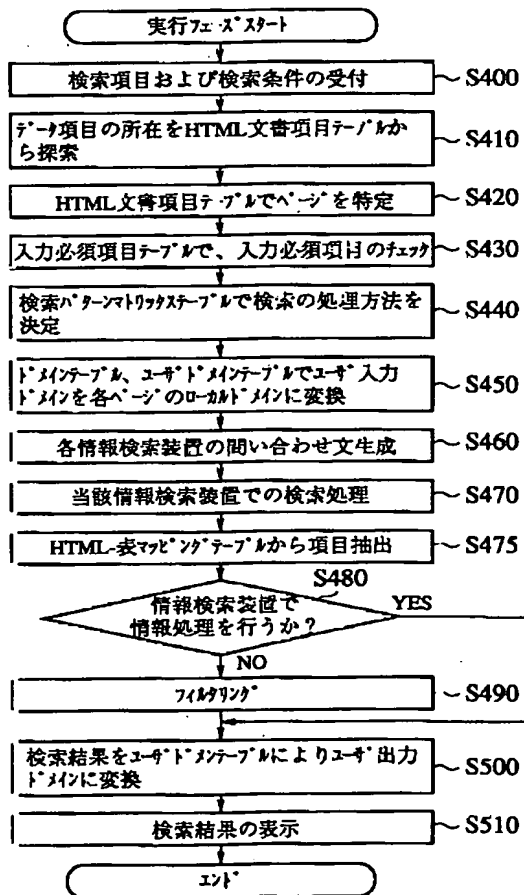
【図25】

ユーザドメインテーブル			
ユーザ名	ドメイングループ	ユーザ入力ドメイン	ユーザ出力ドメイン
ユーザ1	Area	SHITSUKI	SHITSUKI
ユーザ1	Category	RYOURITSUKI	RYOURITSUKI
ユーザ1	Tel	Tel_Paren	Tel_Paren

【図27】



【図29】



【図28】

検索パターンマトリックス							
検索要求		サーチエンジン		処理			
項目	条件	項目	条件	項目		条件	
				検索条件値をそのまま返却	情報源から返却されたものを返却	サーチエンジンで処理	装置でファイナリッシュ
○	○	○	○	×	○	○	×
○	○	○	×	○	×	×	○
○	○	×	○	○	×	○	×
○	×	○	○	×	○	×	×
○	×	○	×	×	○	×	×
○	×	×	○	—	—	—	—
×	○	○	○	×	×	○	×
×	○	○	×	×	×	×	○
×	○	×	○	×	×	○	×

【図31】

検索要求—処理内容テーブル (ページA)							
項目名	検索要求		Page A		項目		
	項目	条件	項目	条件	検索条件値をそのまま返却	情報源から返却されたものを返却	条件
店名	○	×	○	○	×	○	×
電話番号	○	×	○	×	×	○	×
エリア	×	○	○	×	×	×	○
ジャンル	×	○	○	○	×	×	×

【図32】

検索要求—処理内容テーブル (ページB)							
項目名	検索要求		Page B		項目		
	項目	条件	項目	条件	検索条件値をそのまま返却	情報源から返却されたものを返却	条件
店名	○	×	○	○	×	○	×
電話番号	○	×	○	×	×	○	×
エリア	×	○	○	○	×	×	○
ジャンル	×	○	○	×	×	×	○

【図33】

検索要求—処理内容テーブル (ページC)							
項目名	検索要求		Page C		項目		
	項目	条件	項目	条件	検索条件値をそのまま返却	情報源から返却されたものを返却	条件
店名	○	×	○	×	×	○	×
電話番号	○	×	○	×	×	○	×
エリア	×	○	○	○	×	×	○
ジャンル	×	○	○	○	×	×	○



【図35】

Page\_Bの検索結果のHTML

WEBブラウザでの表示例

(A)

店名情報			
店名	エリア	ジャンル	電話番号
B1	横浜市	和食料理	045-****-*****
B2	横浜市	中華料理	045-****-*****
B3	横浜市	中華料理	045-****-*****

HTML文書

(B)

```

<BODY>
<H1>店名情報 </H1>

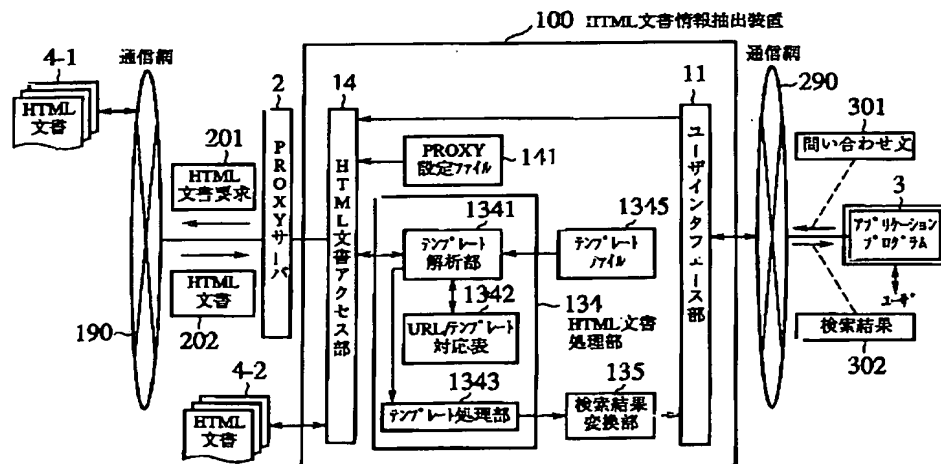
<TABLE BORDER=1>
<TR><TH>店名</TH><TH>エリア</TH><TH>ジャンル</TH><TH>電話番号</TH></TR>
<TR><TD>B1</TD><TD>横浜市</TD><TD>和食料理</TD><TD>045-****-*****</TD></TR>
<TR><TD>B2</TD><TD>横浜市</TD><TD>中華料理</TD><TD>045-****-*****</TD></TR>
<TR><TD>B3</TD><TD>横浜市</TD><TD>中華料理</TD><TD>045-****-*****</TD></TR>
</TABLE>
</BODY>

```

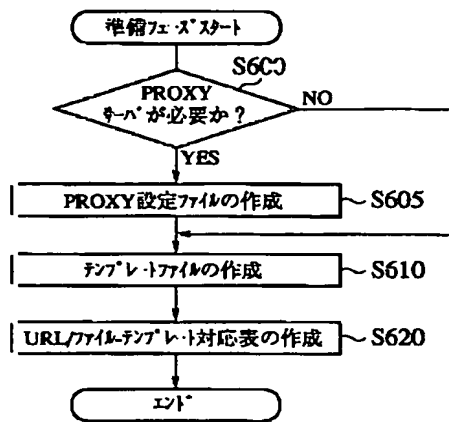
【図36】

ユーザ出力ドメインとロケストメイン			
ページ名	項目	ユーザ出力ドメイン	ロケストメイン
Page_A	電話番号	Tel_Paren	Tel_Paren
Page_B	電話番号	Tel_Paren	Tel_Bar
Page_C	電話番号	Tel_Paren	Tel_Bar

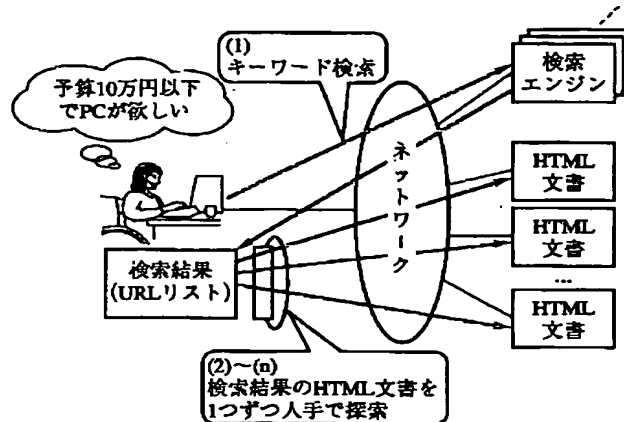
【図37】



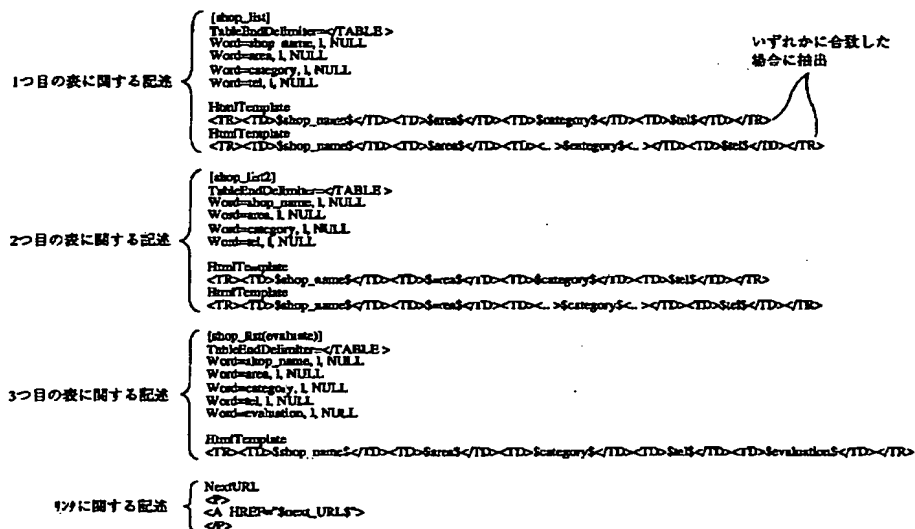
【図38】



【図51】



【図40】



【図48】

1つ目の  
テーブル

店名情報 (その他)				
店名	エリア	ジャンル	電話番号	
飲食店A	鎌倉	中華料理	*****-XYZA	
飲食店X	横須賀	和食料理	*****-YZAB	
飲食店B	逗子	洋食料理	*****-ABCD	
飲食店Z	鎌倉	和食料理	*****-WXYZ	
飲食店Y	逗子	和食料理	*****-QRST	

2つ目の  
テーブル

店名情報 (その他)				
店名	エリア	ジャンル	電話番号	
飲食店C	鎌倉	中華料理	*****-XYZA	
飲食店E	鎌倉	洋食料理	*****-WXYZ	
飲食店D	横須賀	和食料理	*****-YZAB	
飲食店F	逗子	洋食料理	*****-ABCD	
飲食店G	逗子	和食料理	*****-QRST	

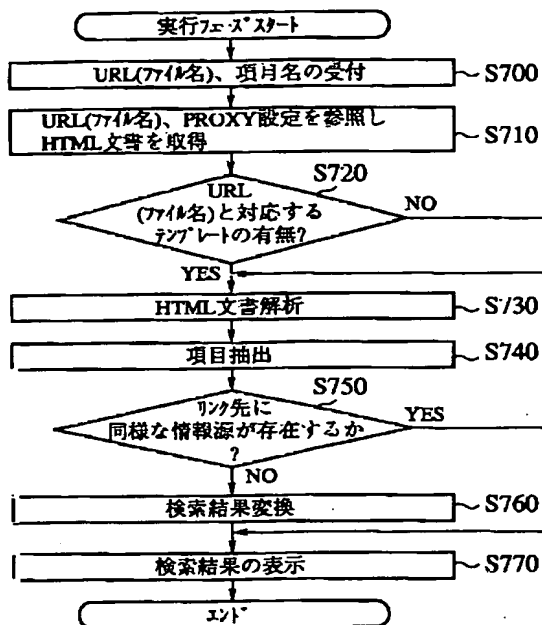
3つ目の  
テーブル

評価あり店名情報 (その他)					
店名	エリア	ジャンル	電話番号	評価	
Q	鎌倉	和食料理	AABB	★★★	
飲食店R	逗子	和食料理	ABAB	★★	
飲食店S	川崎	中華料理	DCAH	★	
飲食店T	鎌倉	洋食料理	WZAB	★★★	

次ページ

次ページ

【図42】



【図45】

Url=<http://www.jba.go.jp/datafile/kachinma/1998east.html>

```
[JBA1998cast]
Word=racename,1,NULL
Word=grade,1,NULL
Word=cercle,1,NULL
Word=round,1,NULL
Word=distance,1,NULL
Word=condition,1,NULL
Word=time,1,NULL
Word=winhorse,1,NULL
Word=sex age,1,NULL
Word=jockey,1,NULL
Word=teki,1,NULL
Word=uri,5,http://www.jb
```

## Html Template

```
<A  
HREF="Suris">"$rcname$"</A><Sgrade$</TD><Scerul$</TD><Smdid$</TD>  
<Sdistanc  
e$</TD><Scondition$</TD><Stime$</TD><..</TD><Swinhouse$</TD><See  
r_age$</TD><  
Sjockey$</TD><Stekis$</TD>  
</TR>
```

【図43】

平成10年重賞競走勝ち馬一覧表 (関東)

選手名	格	場	月日	距離(m)	天候・馬場	141	厩舎馬	性齢	騎手	調教師
小山銀林	C	小山	1.5	2,000	曇・稍重	2:01.4	(外) 砂・砂707417	牡7	岡田 幸雄	田中 昭生
サ・ネリス	C	小山	1.10	9,120	曇・良	R1:59.1	(外) ス・ネ・砂744	牡5	竹 豊	小西 和生
京東杯	C	小山	1.11	1,600	曇・良	1:36.8	(外) 砂・砂7124-	牡4	柴本 善臣	高橋 孝二
7743OC	B	小山	1.25	2,200	晴・良	2:15.3	(外) ジ・砂741	牡5	河田 洋	深見 秀一
ブ・799 S	A	東阪	2.3	9,160	曇・良	1:37.5	(外) 砂・砂707417	牡7	岡田 幸雄	田中 昭生
東阪新開杯	C	東阪	2.8	1,600	曇・良	1:34.2	ス・砂741-	牡5	堤原 正雄	小島 正
東阪障害特別 (春)	-	東阪	2.14	3,300	曇・良	2:39.7	74174174-	牡10	柳 隆彦	森田 人郎
同光通信杯4才S	-	東阪	2.15	9,160	曇・良	1:36.9	(外) コド・スズメ・サ-	牡4	岡島 潤	三ノ宮 敏
74741 S	C	東阪	2.21	3,200	曇・稍重	1:37.6	ネ・1747417	牡6	河田 洋	宇 無 孝幸
女毛C	C	東阪	2.22	1,600	晴・良	1:35.2	(外) ジ・砂7417	牡4	花江 隆徳	須々木 康
小山花馬S	C	小山	3.2	1,800	晴・重	1:49.8	(外) ジ・砂7417417	牡6	立山 典弘	沼江 泰雄
四月賞	B	小山	3.8	2,000	晴・良	2:01.8	74174174-	牡4	竹 豊	黒井 寿仁
五月S	C	小山	3.14	9,180	晴・良	1:51.9	7417417417	牡7	本橋 良喜	古井 一
小山記念	B	小山	3.15	1,800	晴・良	1:48.6	7417417417	牡5	竹 豊	田中 廣
7474-C	C	小山	3.21	1,800	曇・良	1:50.4	(外) エ・砂7417417	牡4	河田 洋	深見 秀一
7474S	B	小山	3.22	1,800	曇・稍重	1:49.8	7417417417	牡4	堤原 正雄	木 貴典
東東賞	C	小山	3.29	2,500	晴・良	2:34.4	ジ・砂7417417	牡9	江原 照男	矢田 昭生
7741S	B	小山	4.5	1,200	晴・良	1:08.5	(外) エ・砂7417417	牡4	飯塚 正新	眞賀 良三
ア・ネ・砂CT	C	小山	4.11	1,600	晴・稍重	1:34.3	(外) 74174174-	牡5	岡田 幸雄	柳 隆彦
小山大障害 (春)	-	小山	4.18	4,100	曇・重	4:46.2	74174174174-	牡9	田山 明	須々木 康
近月賞	A	小山	4.19	2,000	晴・良	2:01.3	7417417417	牡4	立山 典弘	高田 一隆
ジ・7417 T4 S	B	東阪	4.26	1,400	曇・重	1:22.2	(外) コド・スズメ・サ-	牡4	岡島 潤	三ノ宮 敏
8・7417417417	B	東阪	5.2	2,000	晴・良	2:00.3	ス・7417417417	牡4	柳 隆彦	伊 東 敏二
若葉賞	C	東阪	5.9	2,400	曇・良	2:27.6	7417417417	牡4	柴本 善臣	川内 研二
京王杯SC	B	東阪	5.16	1,400	晴・良	R1:20.1	(外) 7417417417	牡5	岡田 幸雄	河原 和雄
NTV杯S	A	東阪	5.17	1,600	晴・稍重	1:33.7	(外) コド・スズメ・サ-	牡4	岡島 潤	三ノ宮 敏
古河大賞典	C	古河	5.17	2,000	曇・良	K1:58.2	7417417417	牡6	吉本 圭	小久保 祥
おゆみ川S	C	東阪	5.23	2,100	晴・良	2:11.8	7417417417	牡7	竹 豊	伊 東 敏二
花の木賞	A	東阪	5.31	9,240	晴・良	2:28.1	ス・砂7417417	牡4	岡島 潤	加 東 敏二
747417417	C	東阪	6.6	1,800	曇・重	1:48.2	(外) エ・砂7417417	牡6	立山 典弘	伊 東 敏二

【 図 4 4 】

```

<HTML>
<HEAD>
<TITLE>平成10年度重賞競争勝ち馬一覧表 (関東) </TITLE>
</HEAD>
<BODY BGCOLOR="#ffffff">
<P>
<DIV ALIGN=CENTER><FONT SIZE=5><B>平成10年度重賞競争勝ち馬一覧表 (関東)
</B></FONT><BF></DIV>
<TABLE BORDER>
<TR>
<TH ALIGN=CENTER BGCOLOR="#808080"NOWRAP>レース名</TH><TH ALIGN=CENTER BGCOLOR=
"#808080"
NOWRAP>格</TH><TH ALIGN=CENTER BGCOLOR="#808080"NOWRAP>場</TH><TH ALIGN=CENTER
BGCOLOR="#808080"NOWRAP>月日</TH><TH ALIGN=CENTER BGCOLOR="#808080"NOWRAP>距離
(m)
<TH><TH ALIGN=CENTER BGCOLOR="#808080"NOWRAP>天候・馬場</TH><TH ALIGN=CENTER
BGCOLOR="#808080"NOWRAP>レース</TH><TH COLSPAN=2 ALIGN=CENTER BGCOLOR="#808080"
NOWRAP>勝ち
馬</TH><TH ALIGN=CENTER BGCOLOR="#808080"NOWRAP>性齢</TH><TH ALIGN=CENTER BGCOL
OR="#808080"
NOWRAP>騎手</TH><TH ALIGN=CENTER BGCOLOR="#808080"NOWRAP>調教師</TH>
</TR>
<TR>
<TD ALIGN=LEFT BGCOLOR="#b5ffb5"><A HREF="1998east/19980105-0611.html">小山銀杯<
/A></TD><TD
ALIGN=CENTER BGCOLOR="#b5ffb5"NOWRAP>C</TD><TD BGCOLOR="#b5ffb5"NOWRAP>小山<
/TD><TD
BGCOLOR="#b5ffb5">1 5</TD><TD ALIGN=RIGHT BGCOLOR="#b5ffb5">2,000</TD>
<TD
BGCOLOR="#b5ffb5">曇・稍重</TD><TD ALIGN=RIGHT BGCOLOR="#b5ffb5">2:01.4</T
D><TD
BGCOLOR="#b5ffb5" NOWRAP>(父)</TD><TD BGCOLOR="#b5ffb5">トウモロコシ</TD>
<TD
BGCOLOR="#b5ffb5">牡7</TD><TD BGCOLOR="#b5ffb5"NOWRAP>岡田 幸雄</TD><TD
BGCOLOR="#b5ffb5">田中 靖孝</TD>
</TR>
<TR>
<TD ALIGN=LEFT BGCOLOR="#b5ffb5"><A HREF="1998east/19980110-0611.html">クワター
15
</A></TD><TD ALIGN=CENTER BGCOLOR="#b5ffb5">C</TD><TD BGCOLOR="#b5ffb5">小山<
/TD><TD
BGCOLOR="#b5ffb5">1.10</TD><TD ALIGN=RIGHT BGCOLOR="#b5ffb5">1,200</T
D><TD
BGCOLOR="#b5ffb5">曇・不良</TD><TD ALIGN=RIGHT BGCOLOR="#b5ffb5">R1:09.1<
/TD><TD
BGCOLOR="#b5ffb5">(外)</TD><TD BGCOLOR="#b5ffb5">スーパードラッグ</TD><TD
BGCOLOR="#b5ffb5">牡5</TD><TD BGCOLOR="#b5ffb5">竹 逸</TD><TD BGCOLOR="#b5ffb5"
>小西 和生
</TD>
</TR>
<TR>
<TD ALIGN=LEFT BGCOLOR="#b5ffb5"><A HREF="1998east/19980111-0611.html">京東杯</A
></TD><TD
ALIGN=CENTER BGCOLOR="#b5ffb5">C</TD><TD BGCOLOR="#b5ffb5">小山</TD><TD BGCOLOR=
"#b5ffb5">

```

【図46】

候補選択番号: 1

騎手	勝ち馬	レース名
岡田 幸雄	メロロンティ	小山銀杯
竹 豊	スーパースター	サニティS
柴木 善臣	ダマシンスキー	京東杯
河田 洋	ジメロノライト	メロノICC
岡田 幸雄	メロロンティ	ブフェラS
蛭原 正義	スーパースター	東阪新聞杯
御堂 信雄	タイタイタイ	東阪障害特別 (春)
岡島羽 均	コンドルエース	同共通信杯4才S
河田 洋	ヒコイッパダン	付添ントS
花沢 隆徳	ダイエタイン	女王C
立山 典弘	ジメロロンダ	小山牝馬S
竹 豊	ウルトライク	四月賞
本橋 広喜	タイロッドアラビ	五月S
竹 豊	ダイキフタス	小山記念
河田 洋	ヒノキエディー	ワラージ
蛭原 正義	クローリケーン	プリンS
江原 照男	ジメロノウガン	東東賞
蛭原 正義	キンギンフェクト	アクリムC
岡田 幸雄	ウルトライク	ブービー御CT
田山 剛	ウルトライク	小山大障害 (春)
立山 典弘	ショウンスカイ	五月賞
岡島羽 均	コンドルエース	ジメロノT4才S
蛭原 正義	スーパースター	スーパースター4才特
柴木 善臣	ダマシンスキー	若葉賞
岡田 幸雄	ダイエタイン	東王杯SC
岡島羽 均	コンドルエース	NTTマイルS
吉本 豊	ウルトライク	占領大賞典
竹 豊	ダイエタイン	おゆみ野S
岡島羽 均	モルフィス	屋の木賞
立山 典弘	ウルトライク	サニティC

【図47】

1つ目の表

店名情報 (横浜)				
店名	エリア	ジャンル	電話番号	
和食店1	横浜	和食料理	***-***-	XXXX
洋食店1	横浜	洋食料理	***-***-	YYYY
和食店2	横浜	和食料理	***-***-	ZZZZ
中華店1	横浜	中華料理	***-***-	WWWW
和食店3	横浜	和食料理	***-***-	QQQQ

2つ目の表

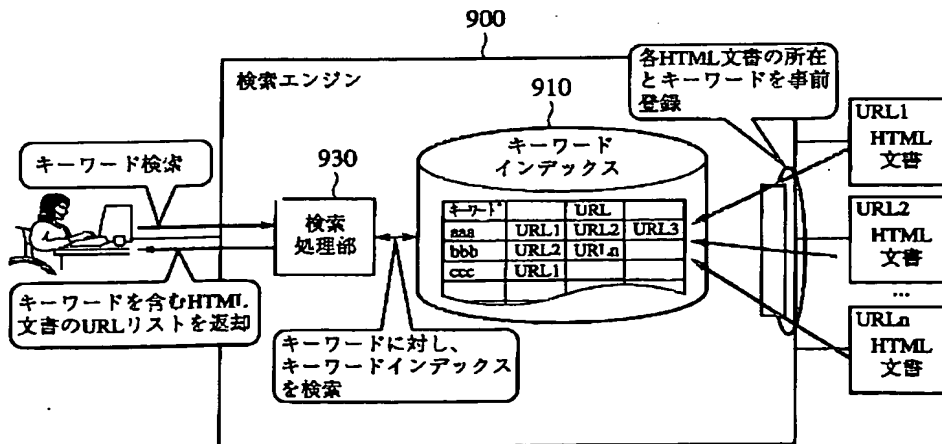
店名情報 (横浜2)				
店名	エリア	ジャンル	電話番号	
洋食店1	横浜	洋食料理	***-***-	SSSS
和食店1	横浜	和食料理	***-***-	RRRR
和食店2	横浜	和食料理	***-***-	TTTT
中華店1	横浜	中華料理	***-***-	UUUU
洋食店2	横浜	洋食料理	***-***-	VVVV

3つ目の表

評価あり店名情報 (横浜)				
店名	エリア	ジャンル	電話番号	評価
和食店1	横浜	和食料理	AAAA	★
洋食店1	横浜	洋食料理	BBBB	★
中華店1	横浜	中華料理	CCCC	★★★
和食店2	横浜	和食料理	DDDD	★★

リンカー 次ページ

【図52】



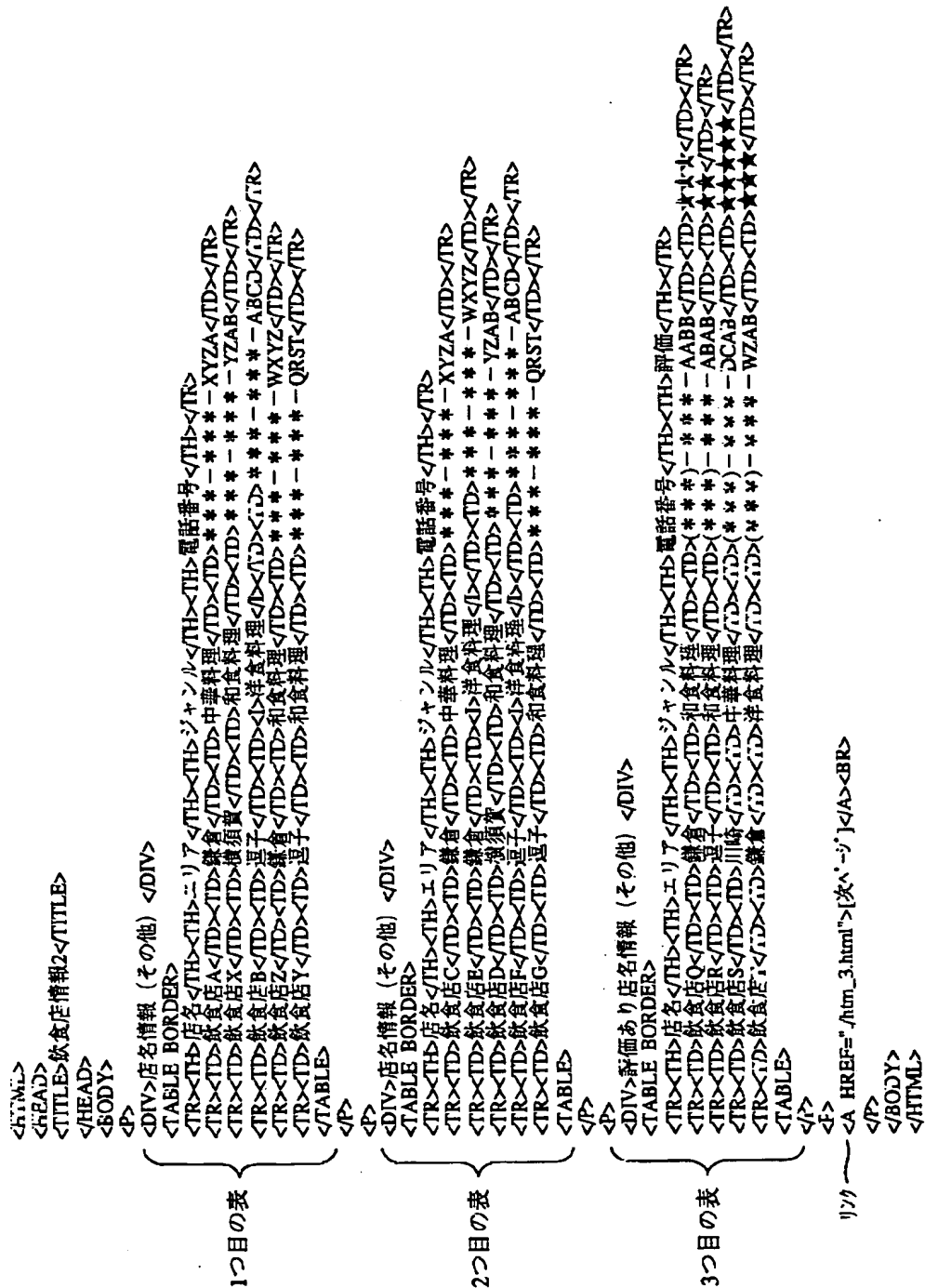
【 図 4 9 】

```

<HTML>
<HEAD>
<TITLE>飲食店情報1</TITLE>
</HEAD>
<BODY>
<P>
<DIV>店名情報 (横浜) </DIV>
<TABLE BORDER>
<TR><TH>店名</TH><TH>エリア</TH><TH>ジャンル</TH><TH>電話番号</TH></TR>
<TR><TD>和食屋1</TD><TD>横浜</TD><TD>和食料理</TD><TD>*****-XXXX</TD></TR>
<TR><TD>洋食屋1</TD><TD>横浜</TD><TD>洋食料理</TD><TD>*****-YYYY</TD></TR>
<TR><TD>和食屋2</TD><TD>横浜</TD><TD>和食料理</TD><TD>*****-ZZZZ</TD></TR>
<TR><TD>中華屋1</TD><TD>横浜</TD><TD>中華料理</TD><TD>*****-WWWW</TD></TR>
<TR><TD>和食屋3</TD><TD>横浜</TD><TD>和食料理</TD><TD>*****-QQQQ</TD></TR>
</TABLE>
</P>
<DIV>店名情報 (横浜2) </DIV>
<TABLE BORDER>
<TR><TH>店名</TH><TH>エリア</TH><TH>ジャンル</TH><TH>電話番号</TH></TR>
<TR><TD>洋食店1</TD><TD>横浜</TD><TD>洋食料理</TD><TD>*****-SSSS</TD></TR>
<TR><TD>和食店1</TD><TD>横浜</TD><TD>和食料理</TD><TD>*****-RRRR</TD></TR>
<TR><TD>和食店2</TD><TD>横浜</TD><TD>和食料理</TD><TD>*****-TTTT</TD></TR>
<TR><TD>中華店1</TD><TD>横浜</TD><TD>中華料理</TD><TD>*****-UUUU</TD></TR>
<TR><TD>洋食店2</TD><TD>横浜</TD><TD>洋食料理</TD><TD>*****-VVVV</TD></TR>
</TABLE>
</P>
<DIV>評価あり店名情報 (横浜) </DIV>
<TABLE BORDER>
<TR><TH>店名</TH><TH>エリア</TH><TH>ジャンル</TH><TH>電話番号</TH><TH>評価</TH></TR>
<TR><TD>和食店1</TD><TD>横浜</TD><TD>和食料理</TD><TD>*****-AAAA</TD><TD>★★★★</TD></TR>
<TR><TD>洋食店1</TD><TD>横浜</TD><TD>洋食料理</TD><TD>*****-BBBB</TD><TD>★★★★</TD></TR>
<TR><TD>中華店1</TD><TD>横浜</TD><TD>中華料理</TD><TD>*****-CCCC</TD><TD>★★★★</TD></TR>
<TR><TD>和食店2</TD><TD>横浜</TD><TD>和食料理</TD><TD>*****-DDDD</TD><TD>★★★★</TD></TR>
</TABLE>
</P>
<P>
<A HREF="/htm_2.html">次へ</A><BR>
</P>
</BODY>
</HTML>

```

【図50】



【手続補正書】

【提出日】平成12年4月25日(2000.4.25)

5)

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】特許請求の範囲

【補正方法】変更

【補正内容】

【特許請求の範囲】

【請求項1】 オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合

検索装置であって、  
 オープンネットワーク上での半構造化文書の所在を示す  
 所在情報と、前記半構造化文書の構造を、抽出すべき項  
 目ごとに区切るための文書構造情報と、前記項目ごとに  
 前記項目を条件検索するために用いるデータ属性を規定  
 するデータ属性情報と、ユーザーの表示における項目の  
 表現形式、各半構造化文書の項目の表現形式およびこれ  
 らの間の表現形式を変換するために用いる関数を定義す  
 る表現形式変換情報とを、各半構造化文書の項目情報を  
 記述するために参照されるメタデータとして記憶する記  
 憶部と、  
 検索項目および検索条件からなる入力された問い合わせ  
 に基づいて、すべての検索項目に対応する項目を有する  
 半構造化文書の所在を前記所在情報から得る文書所在探  
 索部と、  
 入力された前記問い合わせを、前記表現形式変換情報に  
 基づいて、前記得られた所在にある半構造化文書中の前  
 記検索項目に対応する項目の表現形式に必要な応じ前記  
 関数を参照して変換する問い合わせ変換部と、  
 前記変換された問い合わせを前記得られた所在に送信し  
 て、半構造化文書を取得する文書検索部と、  
 取得された各半構造化文書から、前記文書構造情報に基  
 づいて、項目データを抽出し、前記検索条件を用い、前  
 記データ属性情報に基づいて前記抽出された項目データ  
 を選択して検索結果とする文書処理部と、  
 前記検索結果を、前記表現形式変換情報に基づいて、前  
 記検索結果中の項目に対応する各ユーザーごとに定義さ  
 れた項目の表現形式に必要な応じ前記関数を参照して変  
 換する検索結果変換部とを具備することを特徴とする半  
 構造化文書情報統合検索装置。  
 【請求項2】 上記半構造化文書情報統合検索装置は、  
 さらに、  
 半構造化文書ごとに、半構造化文書の構造を抽出すべき  
 項目ごとに区切るための文書構造情報に基づき、少なく  
 とも抽出すべき項目名と、半構造化文書から抽出すべき  
 項目群の抽出テキスト形式情報を記述するテンプレート  
 を記憶するテンプレート記憶部と、  
 取得された半構造化文書に対応するテンプレートを解析  
 するテンプレート解析部と、  
 前記取得された半構造化文書をスキャンして、該半構造  
 化文書と、前記テンプレートとを比較して、前記抽出テ  
 キスト形式情報に合致した項目の項目データを抽出する  
 テンプレート処理部とを具備し、  
 前記テンプレートには、各項目データに対応する変数名  
 が記述されるとともに、半構造化文書が複数の部分構造  
 を有する場合、各部分構造ごとに抽出テキスト形式情報  
 が記述され、  
 前記テンプレート処理部は、抽出された項目データを、  
 各部分構造ごとの検索結果とすることを特徴とする請求  
 項1に記載の半構造化文書情報統合検索装置。

【請求項3】 前記文書処理部は、前記検索結果を、表  
 形式に整形することを特徴とする請求項1に記載の半構  
 造化文書情報統合検索装置。

【請求項4】 前記文書処理部は、前記テンプレート中  
 の前記抽出テキスト形式情報が、他の半構造化文書への  
 リンク情報を含む場合には、リンク先の半構造化文書を  
 スキャンして、前記リンク先の半構造化文書と前記テン  
 プレートとを比較することを特徴とする請求項2に記載  
 の半構造化文書情報統合検索装置。

【請求項5】 前記テンプレートは、半構造化文書の各  
 部分構造に対して、前記部分構造の一部に存在する、前  
 記文書構造情報が他の部分と異なる部分をそれぞれ抽出  
 するための、異なるタグにそれぞれ対応する複数の抽出  
 テキスト形式情報が記述され、

前記テンプレート処理部は、前記取得された第1の検索  
 結果である半構造化文書をスキャンして、該半構造化文  
 書と、該半構造化文書に対応する前記テンプレート中の  
 前記複数の抽出テキスト形式情報のいずれかが合致した  
 場合に、合致した項目の項目データを抽出することを特  
 徴とする請求項2に記載の半構造化文書情報統合検索装  
 置。

【請求項6】 前記テンプレートは、半構造化文書が互  
 いに異なる要素からなる複数の部分構造を有する場合、  
 各部分構造ごとに抽出テキスト形式情報が記述され、  
 前記テンプレート処理部は、抽出された項目データを、  
 各部分構造ごとの検索結果とすることを特徴とする請求  
 項2に記載の半構造化文書情報統合検索装置。

【請求項7】 オープンネットワーク上の複数のサーチ  
 エンジンを経由して情報を検索する半構造化文書情報統  
 合検索装置であって、

オープンネットワーク上でのサーチエンジンの所在を示  
 す所在情報と、各サーチエンジンへの入力フォームにお  
 いて入力が必要とされる入力必須項目を定義する入力必  
 須項目情報と、HTML文書の構造を、抽出すべき項目  
 ごとに区切るための文書構造情報と、項目ごとに各サー  
 チエンジン内において該項目が取得可能か否かおよび条  
 件指定可能か否かを示す項目属性情報と、前記項目ごと  
 に前記項目を条件検索するためのデータ属性を規定する  
 データ属性情報と、ユーザーの項目の表現形式と各HTML  
 文書の項目の表現形式との間の変換情報を定義する  
 表現形式変換情報とを記憶する記憶部と、

検索項目および検索条件からなるユーザーから入力され  
 た問い合わせに基づいて、すべての検索項目に対応する  
 項目を有するサーチエンジンの所在を前記所在情報から  
 得る文書所在探索部と、

前記入力必須項目情報に基づいて、各サーチエンジンに  
 おける入力必須項目と前記入力された問い合わせで指定  
 された項目とを比較することにより、前記得られた所在  
 のサーチエンジンの中から、前記入力必須項目を満たす  
 検索項目が指定されたサーチエンジンを、検索対象サー



チエンジンとして選択するサーチエンジン選択部と、前記入力された検索項目および検索条件と、前記項目属性情報との組み合わせを規定するマトリックステーブルに基づき各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定部と、前記変換された問い合わせ群のそれぞれを、前記表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換部と、前記変換された問い合わせを前記得られた所在に送信して、HTML文書を取得する文書検索部と、各サーチエンジンから取得されたHTML文書からなる第1の検索結果から、前記文書構造情報に基づいて、項目データを抽出するとともに、少なくともサーチエンジンにおいて条件検索が実行されなかった項目に関し、対応する前記検索処理パターンに従い、前記検索条件および前記データ属性情報に基づいて、抽出された項目データから前記検索条件に合致する項目データを選択して、第2の検索結果とする検索結果処理部と、前記第2の検索結果を、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置。

【請求項8】 上記半構造化文書情報統合検索装置は、さらに、HTML文書ごとに、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部と、取得されたHTML文書に対応するテンプレートを解析するテンプレート解析部と、前記取得されたHTML文書をスキャンして、該HTML文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理部とを具備し、前記テンプレートには、各項目データに対応する変数名が記述されるとともに、HTML文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項7に記載の半構造化文書情報統合検索装置。

【請求項9】 前記文書処理部は、前記検索結果を、表形式に整形することを特徴とする請求項7に記載の半構造化文書情報統合検索装置。

【請求項10】 前記文書処理部は、前記テンプレート

中の前記抽出テキスト形式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較することを特徴とする請求項8に記載の半構造化文書情報統合検索装置。

【請求項11】 前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する、前記文書構造情報が他の部分と異なる部分をそれぞれ抽出するための、異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、前記テンプレート処理部は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項8に記載の半構造化文書情報統合検索装置。

【請求項12】 前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項8に記載の半構造化文書情報統合検索装置。

【請求項13】 オープンネットワーク上の任意のHTML文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出装置であって、HTML文書ごとに、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部と、取得されたHTML文書に対応するテンプレートを解析するテンプレート解析部と、前記取得されたHTML文書をスキャンして、該HTML文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理部とを具備し、前記テンプレートには、各項目データに対応する変数名が記述されるとともに、HTML文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする半構造化文書情報抽出装置。

【請求項14】 前記テンプレート処理部は、前記抽出された項目データを、表形式に整形することを特徴とする請求項13に記載の半構造化文書情報抽出装置。

【請求項15】 前記テンプレート処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のHTML

L文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較することを特徴とする請求項13に記載の半構造化文書情報抽出装置。

【請求項16】 前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する、前記文書構造情報が他の部分と異なる部分をそれぞれ抽出するための、異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、

前記テンプレート処理部は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項13に記載の半構造化文書情報抽出装置。

【請求項17】 前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項13に記載の半構造化文書情報抽出装置。

【請求項18】 オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索方法であって、

オープンネットワーク上で半構造化文書の所在を示す所在情報と、前記半構造化文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するために用いるデータ属性を規定するデータ属性情報と、ユーザーの表示における項目の表現形式、各半構造化文書の項目の表現形式およびこれらの間の表現形式を変換するために用いる関数を定義する表現形式変換情報とを、各半構造化文書の項目情報を記述するために参照されるメタデータとして記憶するステップと、

検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を前記所在情報から得るステップと、

入力された前記問い合わせを、前記表現形式変換情報に基づいて、前記得られた所在にある半構造化文書中の前記検索項目に対応する項目の表現形式に必要に応じ前記関数を参照して変換するステップと、

前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得するステップと、

取得された各半構造化文書から、前記文書構造情報に基づいて、項目データを抽出し、前記検索条件を用い、前記データ属性情報に基づいて前記抽出された項目データを選択して検索結果とするステップと、

前記検索結果を、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に必要に応じ前記関数を参照して変換するステップとを含むことを特徴とする半構造化文書情報統合検索方法。

【請求項19】 オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索方法であって、

オープンネットワーク上でサーチエンジンの所在を示す所在情報と、各サーチエンジンへの入力フォームにおいて入力が必要とされる入力必須項目を定義する入力必須項目情報と、HTML文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、項目ごとに各サーチエンジン内において該項目が取得可能か否かおよび条件指定可能か否かを示す項目属性情報と、前記項目ごとに前記項目を条件検索するためのデータ属性を規定するデータ属性情報と、ユーザーの項目の表現形式と各HTML文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶するステップと、

検索項目および検索条件からなるユーザーから入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を前記所在情報から得るステップと、

前記入力必須項目情報に基づいて、各サーチエンジンにおける入力必須項目と前記入力された問い合わせで指定された項目とを比較することにより、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索項目が指定されたサーチエンジンを、検索対象サーチエンジンとして選択するステップと、

前記入力された検索項目および検索条件と、前記項目属性情報との組み合わせを規定するマトリックステーブルに基づき各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換するステップと、

前記変換された問い合わせ群のそれぞれを、前記表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換するステップと、

前記変換された問い合わせを前記得られた所在に送信して、HTML文書を取得するステップと、

各サーチエンジンから取得されたHTML文書からなる第1の検索結果から、前記文書構造情報に基づいて、項目データを抽出するとともに、少なくともサーチエンジンにおいて条件検索が実行されなかった項目に関し、対応する前記検索処理パターンに従い、前記検索条件および前記データ属性情報に基づいて、抽出された項目データから前記検索条件に合致する項目データを選択して、第2の検索結果とするステップと、

前記第2の検索結果を、前記表現形式変換情報に基づい

て、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換するステップとを含むことを特徴とする半構造化文書情報統合検索方法。

【請求項 2 0】 オープンネットワーク上の任意の HTML 文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出方法であって、HTML 文書ごとに、HTML 文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML 文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するステップと、

取得された HTML 文書に対応するテンプレートを解析するステップと、

前記取得された HTML 文書をスキャンして、該 HTML 文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するステップとを含む、

前記テンプレートには、各項目データに対応する変数名が記述されるとともに、HTML 文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記項目データを抽出するステップは、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする半構造化文書情報抽出方法。

【請求項 2 1】 オープンネットワーク上の複数の半構造化文書に内在する情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、

オープンネットワーク上での半構造化文書の所在を示す所在情報と、前記半構造化文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するために用いるデータ属性を規定するデータ属性情報と、ユーザーの表示における項目の表現形式、各半構造化文書の項目の表現形式およびこれらの間の表現形式を変換するために用いる関数を定義する表現形式変換情報とを、各半構造化文書の項目情報を記述するために参照されるメタデータとして記憶する記憶処理と、

検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を前記所在情報から得る文書所在探索処理と、

入力された前記問い合わせを、前記表現形式変換情報に基づいて、前記得られた所在にある半構造化文書中の前記検索項目に対応する項目の表現形式に必要な応じ前記関数を参照して変換する問い合わせ変換処理と、

前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索処理と、

取得された各半構造化文書から、前記文書構造情報に基づいて、項目データを抽出し、前記検索条件を用い、前

記データ属性情報に基づいて前記抽出された項目データを選択して検索結果とする文書処理と、

前記検索結果を、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に必要な応じ前記関数を参照して変換する検索結果変換処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項 2 2】 上記コンピュータ読み取り可能な記録媒体は、さらに、

半構造化文書ごとに、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶処理と、

取得された半構造化文書に対応するテンプレートを解析するテンプレート解析処理と、

前記取得された半構造化文書をスキャンして、該半構造化文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理とを含む、

前記テンプレートには、各項目データに対応する変数名が記述されるとともに、半構造化文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項 2 1 に記載のコンピュータ読み取り可能な記録媒体。

【請求項 2 3】 前記文書処理は、前記検索結果を、表形式に整形することを特徴とする請求項 2 1 に記載のコンピュータ読み取り可能な記録媒体。

【請求項 2 4】 前記文書処理は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較することを特徴とする請求項 2 2 に記載のコンピュータ読み取り可能な記録媒体。

【請求項 2 5】 前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する、前記文書構造情報が他の部分と異なる部分をそれぞれ抽出するための、異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、

前記テンプレート処理は、前記取得された第 1 の検索結果である半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項 2 2 に記載のコンピュータ読み取り可能な記録媒体。

【請求項 2 6】 前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場

合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項22に記載のコンピュータ読み取り可能な記録媒体。

【請求項27】 オープンネットワーク上の複数のサーチエンジンを介して情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、

オープンネットワーク上で各サーチエンジンの所在を示す所在情報と、各サーチエンジンへの入力フォームにおいて入力が必要とされる入力必須項目を定義する入力必須項目情報と、HTML文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、項目ごとに各サーチエンジン内において該項目が取得可能か否かおよび条件指定可能か否かを示す項目属性情報と、前記項目ごとに前記項目を条件検索するためのデータ属性を規定するデータ属性情報と、ユーザーの項目の表現形式と各HTML文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶処理と、

検索項目および検索条件からなるユーザーから入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を前記所在情報から得る文書所在探索処理と、

前記入力必須項目情報に基づいて、各サーチエンジンにおける入力必須項目と前記入力された問い合わせで指定された項目とを比較することにより、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索項目が指定されたサーチエンジンを、検索対象サーチエンジンとして選択するサーチエンジン選択処理と、前記入力された検索項目および検索条件と、前記項目属性情報との組み合わせを規定するマトリックステーブルに基づき各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定処理と、

前記変換された問い合わせ群のそれぞれを、前記表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換処理と、

前記変換された問い合わせを前記得られた所在に送信して、HTML文書を取得する文書検索処理と、各サーチエンジンから取得されたHTML文書からなる第1の検索結果から、前記文書構造情報に基づいて、項目データを抽出するとともに、少なくともサーチエンジン内において条件検索が実行されなかった項目に関し、対応する前記検索処理パターンに従い、前記検索条件および前記属性情報に基づいて、抽出された項目データから前記検索条件に合致する項目データを選択して、第2の検索結果とする検索結果生成処理と、

前記第2の検索結果を、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換処理とを具備することを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項28】 上記コンピュータ読み取り可能な記録媒体は、さらに、

HTML文書ごとに、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶処理と、

取得されたHTML文書に対応するテンプレートを解析するテンプレート解析処理と、

前記取得されたHTML文書をスキャンして、該HTML文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理とを含み、

前記テンプレートには、各項目データに対応する変数名が記述されるとともに、HTML文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項27に記載のコンピュータ読み取り可能な記録媒体。

【請求項29】 前記文書処理は、前記検索結果を、表形式に整形することを特徴とする請求項27に記載のコンピュータ読み取り可能な記録媒体。

【請求項30】 前記文書処理は、前記テンプレート中の前記抽出テキスト形式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較することを特徴とする請求項28に記載のコンピュータ読み取り可能な記録媒体。

【請求項31】 前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する、前記部分構造情報が異なる部分をそれぞれ抽出するための異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、

前記テンプレート処理は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項28に記載のコンピュータ読み取り可能な記録媒体。

【請求項32】 前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項28に記載のコンピュータ読み取り可能な記録媒体。

【請求項33】 オープンネットワーク上の任意のHTML文書に内在する情報から項目ごとに情報を抽出する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、HTML文書ごとに、HTML文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、HTML文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶処理と、取得されたHTML文書に対応するテンプレートを解析するテンプレート解析処理と、前記取得されたHTML文書をスキャンして、該HTML文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理とを含み、前記テンプレートには、各項目データに対応する変数名が記述されるとともに、HTML文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項34】 前記テンプレート処理は、前記抽出された項目データを、表形式に整形することを特徴とする

請求項33に記載のコンピュータ読み取り可能な記録媒体。

【請求項35】 前記テンプレート処理は、前記テンプレート中の前記抽出テキスト形式情報が、他のHTML文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較することを特徴とする請求項33に記載のコンピュータ読み取り可能な記録媒体。

【請求項36】 前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する、前記文書構造情報が他の部分と異なる部分をそれぞれ抽出するための、異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、前記テンプレート処理は、前記取得された第1の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項33に記載のコンピュータ読み取り可能な記録媒体。

【請求項37】 前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項33に記載のコンピュータ読み取り可能な記録媒体。

#### フロントページの続き

(72)発明者 永末 壽宏  
東京都新宿区西新宿三丁目19番2号 日本  
電信電話株式会社内

(72)発明者 星野 隆  
東京都新宿区西新宿三丁目19番2号 日本  
電信電話株式会社内

(72)発明者 町原 宏毅  
東京都新宿区西新宿三丁目19番2号 日本  
電信電話株式会社内

Fターム(参考) 5B075 KK02 KK03 KK07 ND34 QM05  
5B082 GA02 GC04